PONTIFICIA UNIVERSIDAD CATÓLICA DEL PERÚ

ESCUELA DE POSGRADO

MODELOS DE REGRESIÓN GAMMA GENERALIZADA CERO-INFLACIONADA PARA LA MEDIA CON APLICACIÓN A GASTOS EN EDUCACIÓN

TESIS PARA OPTAR EL GRADO DE
MAGÍSTER EN ESTADÍSTICA

Presentado por:

Aníbal Alcides Vásquez Beltrán

Asesor:
Dr. Luis Hilmar Valdivieso Serrano

Miembros del jurado:
Dr. Luis Hilmar Valdivieso Serrano
Dr. Cristian Luis Bayes Rodríguez
Dr. Víctor Giancarlo Sal y Rosas Celi

Lima, 2018

Resumen

Cuando los valores posibles de una variable aleatoria son continuos y no negativos, incluyendo el valor cero con probabilidad no nula, la variable es denominada semicontinua o cero-inflacionada y posiblemente sea pertinente suponer que presenta una distribución mixta de probabilidades constituida por una distribución de Bernoulli para explicar si la respuesta toma el valor cero o no y una distribución continua positiva para explicar si ésta última no es cero. En el análisis de regresión, el modelo de dos partes (MDP) es tradicionalmente usado para explicar una variable semicontinua. En el MDP la respuesta presenta este tipo de distribución mixta y sus parámetros son expresados de tal manera que posibilite estimar el efecto de un conjunto de covariables sobre la media de esta respuesta condicionada a que tome valores positivos y sobre la probabilidad de que la respuesta tome el valor cero.

El objetivo de la tesis es estudiar un modelo alternativo al MDP, que llamaremos modelo de regresión cero-inflacionada a la media (MCIM), cuya parametrización permita estimar e interpretar efectos de covariables sobre la media total de la respuesta, en lugar de la media condicionada a valores positivos. Además, optamos por la distribución gamma generalizada (MCIM-GG) para modelar ciertas características de los valores positivos de la respuesta, tales como, por ejemplo, la asimetría positiva y la curtosis pronunciada. Estas características, junto con el exceso de valores cero, son típicas en diferentes ejemplos de variables respuestas en la Economía y la Medicina.

Los resultados del estudio de simulación muestran un adecuado desempeño de las estimaciones de máxima verosimilitud del MCIM-GG bajo diferentes escenarios definidos según porcentajes de valores ceros de la respuesta y tamaños de muestra. Por último, los resultados de la aplicación muestran que el MCIM-GG puede tener un mejor ajuste a los datos respecto al MDP-GG, así como proporcionar una más directa interpretación de los efectos de ciertas covariables sobre la media de los gastos en educación de adolescentes participantes del estudio Niños del Milenio en el Perú.

Palabras claves: variable semicontinua o cero-inflacionada, distribución gamma generalizada, regresión de dos partes, regresión cero-inflacionada, estimación de máxima verosimilitud.

Índice general

Índice de figuras

Índice de cuadros

Capítulo 1

Introducción

1.1. Consideraciones preliminares

En diversas investigaciones se presenta la necesidad de explicar o predecir una variable cuyos posibles valores son continuos y no negativos, incluyendo el valor cero con probabilidad no nula. Variables como estas suelen ser denominadas semicontinuas, cero-inflacionadas o mixturas con masa en el punto cero.

En diversas investigaciones en Economía y Medicina encontraremos ejemplos de este tipo de variable. En un estudio de gastos en servicios médicos de un plan de seguro, es posible que algunos asegurados no usen los servicios y no efectúen gastos en un período de disponibilidad del seguro (Duan et al., 1983). En un estudio de montos de pérdida de bancos generados por el incumplimiento de préstamos de sus clientes, es posible que parte de los clientes no presenten préstamos incumplidos en el momento del estudio (Tong et al., 2013). En un estudio sobre niveles de daño al funcionamiento físico de pacientes de alguna enfermedad, se utiliza una medición que toma el valor de cero si un paciente no presenta tal daño o que toma un rango de valores continuos para señalar el nivel del daño (Su et al., 2009). Si bien son ejemplos de variables que toman valores continuos positivos, está la posibilidad que también tomen el valor cero, inclusive en exceso o con elevada frecuencia.

El análisis de regresión de una variable respuesta semicontinua se ha realizado tradicionalmente mediante el modelo de dos partes (MDP), propuesto por Duan et al. (1983). Con este modelo se asume que la variable respuesta Y presenta una distribución mixta de probabilidad conformada por una distribución de Bernoulli para explicar si toma el valor cero o no y una distribución continua positiva para cuando esta última no es cero. Además, el modelo plantea ecuaciones de regresión para explicar la probabilidad de que la respuesta sea cero y la media de la respuesta condicionada a valores positivos.

Esta tesis estudia un modelo alternativo al MDP, que denominaremos modelo de regresión cero-inflacionada a la media (MCIM), cuya parametrización permite estimar en un solo proceso de optimización los efectos de un conjunto de covariables sobre la media total de la respuesta, $E(Y)$, en vez de la media condicionada, $E(Y|Y > 0)$. Esta parametrización es similar a la propuesta en Bayes y Valdivieso (2016) para respuestas acotadas al intervalo

cerrado $[0, 1]$ y en Smith et al. (2014) para respuestas semicontinuas. Asimismo, planteamos otra ecuación de regresión para estimar los efectos sobre la probabilidad de que la respuesta sea cero. Estudiaremos el MCIM suponiendo una distribución que ayude a modelar la asimetría positiva y la curtosis pronunciada de los valores positivos de la respuesta. Elegiremos la distribución gamma generalizada (MCIM-GG), que incluye como casos particulares a las distribuciones gamma estándar (MCIM-G) y log-normal (MCIM-LN), entre otras.

Mediante la simulación de escenarios con diferentes porcentajes de valores cero en la respuesta y tamaños de muestra, evaluaremos el desempeño de las estimaciones de máxima verosimilitud del MCIM-GG en términos de sesgo, sesgo porcentual y raíz del error cuadrático medio. Finalmente, el modelo será aplicado para estudiar los determinantes de los niveles de gasto destinado a educación sobre la base del estudio Niños del Milenio (*Young Lives*) del año 2009 en el Perú.

1.2. Objetivos de la tesis

El objetivo general de la tesis es estudiar las características del MCIM.

De manera específica:

- Revisar la literatura acerca de los modelos MDP y MCIM para explicar una variable semicontinua, identificando sus ventajas y limitaciones.
- Estudiar las propiedades e implementar la estimación por máxima verosimilitud del MCIM utilizando la distribución gamma generalizada.
- Realizar simulaciones de diferentes escenarios del MCIM y así evaluar el desempeño de sus estimadores.
- Aplicar y comparar los modelos MCIM y MDP con datos del estudio Niños del Milenio del año 2009 en el Perú.

1.3. Organización de la tesis

La tesis está dividida en 6 capítulos. En el Capítulo 2 presentamos los conceptos básicos necesarios para la comprensión del MCIM. En el Capítulo 3 presentamos el desarrollo del MCIM y el método de estimación de los parámetros de regresión. En el Capítulo 4 realizamos un estudio de simulación del MCIM bajo diferentes escenarios. En el Capítulo 5 desarrollamos la aplicación del MCIM a datos reales y comparamos sus resultados con los de la aplicación del MDP. Por último, en el Capítulo 6 presentamos las conclusiones de la tesis y las sugerencias de investigaciones futuras.

Capítulo 2

Conceptos básicos

En este capítulo presentaremos los conceptos básicos necesarios para una mejor comprensión de los modelos de regresión que comúnmente son utilizados para explicar una variable semicontinua.

2.1. La distribución gamma

Es una distribución de probabilidades aplicable a variables continuas y positivas, de frecuente uso para modelar variables con asimetría positiva y cola relativamente ligera.

Una variable aleatoria continua Y tiene una distribución gamma de parámetros $\alpha > 0$ y $\beta > 0$, si su función de densidad está dada por:

$$g(y \,|\, \alpha, \beta) = \frac{1}{\Gamma(\alpha)} \left(\frac{1}{\beta}\right)^{\alpha} y^{\alpha-1} \exp\left(-\frac{y}{\beta}\right), \qquad y > 0, \tag{2.1}$$

donde α es un parámetro de forma, β es un parámetro de escala y $\Gamma(\alpha)$ es la función gamma definida como $\Gamma(\alpha) = \int_0^{\infty} y^{\alpha-1} \exp(-y) dy$.

El parámetro α tiene influencia en el apuntalamiento de la densidad, y el parámetro β tiene influencia en la extensión del extremo derecho de la distribución. La media y varianza de esta distribución están dadas respectivamente por:

$$E(Y) = \alpha\beta \tag{2.2}$$

$$V(Y) = \alpha\beta^2 \tag{2.3}$$

Con la finalidad de simplificar la interpretación de los parámetros, en el sentido de interpretarlos en términos de la variable respuesta en su escala original, comúnmente se reparametriza la distribución de tal forma que esté explícitamente en función de la media de la variable; por ello, consideramos el parámetro μ dado por:

$$\mu = \alpha\beta \tag{2.4}$$

Ahora una variable aleatoria Y tiene distribución gamma en términos de μ y α si su función de densidad está dada por:

$$g(y \,|\, \mu, \alpha) = \frac{1}{\Gamma(\alpha)} \left(\frac{\alpha}{\mu}\right)^{\alpha} y^{\alpha-1} \exp\left(\frac{-y\alpha}{\mu}\right), \qquad y > 0. \tag{2.5}$$

La media y varianza de esta distribución están dadas respectivamente por:

$$E(Y) = \mu \tag{2.6}$$

$$V(Y) = \frac{\mu^2}{\alpha} \tag{2.7}$$

Note que α puede ser considerado también como un parámetro de precisión, porque, para un valor dado del parámetro μ, la varianza de esta distribución disminuye conforme el parámetro α aumenta. Además, de acuerdo a los coeficientes de Fisher de asimetría, $A(Y)$, y de curtosis, $K(Y)$, se verifica que α tiene un efecto negativo sobre la asimetría y la curtosis de esta distribución:

$$A(Y) = \frac{E\left((Y - E(Y))^3\right)}{V(Y)^{3/2}} = \frac{2}{\sqrt{\alpha}} \tag{2.8}$$

$$K(Y) = \frac{E\left((Y - E(Y))^4\right)}{V(Y)^2} = 3 + \frac{6}{\alpha} \tag{2.9}$$

2.2. La distribución gamma cero-inflacionada

Una distribución de probabilidades que pueda modelar una variable continua y no negativa, que incluya al valor cero con probabilidad no nula, es la distribución cero-inflacionada. En ella se asume que la variable tiene una distribución mixta de probabilidad: una distribución de Bernoulli para explicar si toma el valor cero o no y una distribución continua positiva para cuando esta última no es cero.

Una variable aleatoria Y se dice que presenta una distribución gamma cero-inflacionada de parámetros $0 < \delta < 1$, $\mu > 0$ y $\alpha > 0$ si su función de masa de probabilidad está dada por:

$$f(y \,|\, \delta, \mu, \alpha) = \begin{cases} \delta & , \text{ si } y = 0 \\ (1-\delta)\, g(y \,|\, \mu, \alpha) & , \text{ si } y > 0 \end{cases} \tag{2.10}$$

donde el parámetro δ es la probabilidad de que la variable tome el valor cero y la función $g(y \,|\, \alpha, \mu)$ es la función de densidad de la distribución gamma en términos de (2.5). La media y varianza de esta distribución cero-inflacionada están dadas respectivamente por:

$$E(Y) = (1-\delta)\,\mu \tag{2.11}$$

$$V(Y) = \mu^2(1-\delta)\left(\frac{1+\alpha\delta}{\alpha}\right) \tag{2.12}$$

El análisis tradicional de regresión para este modelo plantea ecuaciones de regresión para modelar la probabilidad de que la variable aleatoria tome el valor cero, $\delta = P(Y=0)$, y la media de la variable condicionada a valores positivos, $\mu = E(Y|Y>0)$. Sin embargo, el énfasis de alguna investigación puede estar centrada en estimar la media total de la variable, $E(Y)$, en vez de la media de solamente los valores positivos. Esto hace necesaria la reparametrización de la distribución de tal forma que sea posible estimar directamente la media total. Por ello, se considera un nuevo parámetro γ igual a la media de la variable gamma cero-inflacionada. Ésta es dada por:

$$\gamma = (1-\delta)\,\mu \tag{2.13}$$

La función de densidad de la distribución gamma en términos de esta nueva parametrización está ahora dada por:

$$g(y\,|\,\delta,\gamma,\alpha) = \frac{1}{\Gamma(\alpha)}\left[\frac{(1-\delta)\alpha}{\gamma}\right]^{\alpha} y^{\alpha-1}\,\exp\left(-y\frac{(1-\delta)\alpha}{\gamma}\right), \qquad y>0 \tag{2.14}$$

Por consiguiente, una variable aleatoria Y presenta distribución gamma cero-inflacionada de parámetros $0<\delta<1$, $\gamma>0$ y $\alpha>0$ si su función de masa de probabilidad está dada por:

$$f(y\,|\,\delta,\gamma,\alpha) = \begin{cases} \delta & ,\ \text{si } y=0 \\ (1-\delta)\,g(y\,|\,\delta,\gamma,\alpha) & ,\ \text{si } y>0 \end{cases} \tag{2.15}$$

La media y varianza de esta distribución están dadas respectivamente por:

$$E(Y) = \gamma \tag{2.16}$$

$$V(Y) = \gamma^2\left(\frac{\frac{1}{\alpha}+\delta}{1-\delta}\right) \tag{2.17}$$

Como era esperable, la Figura 2.1 muestra que la función de densidad de una variable gamma cero-inflacionada en términos de (2.15) es continua excepto cuando la variable toma el valor cero. La asimetría positiva de la distribución, que es una especial característica de la distribución gamma, es más acentuada cuando el parámetro α disminuye. Asimismo, cuando el parámetro γ disminuye y $\alpha>1$, la distribución tiene mayor apuntalamiento, concentrándose las observaciones cerca al valor de γ.

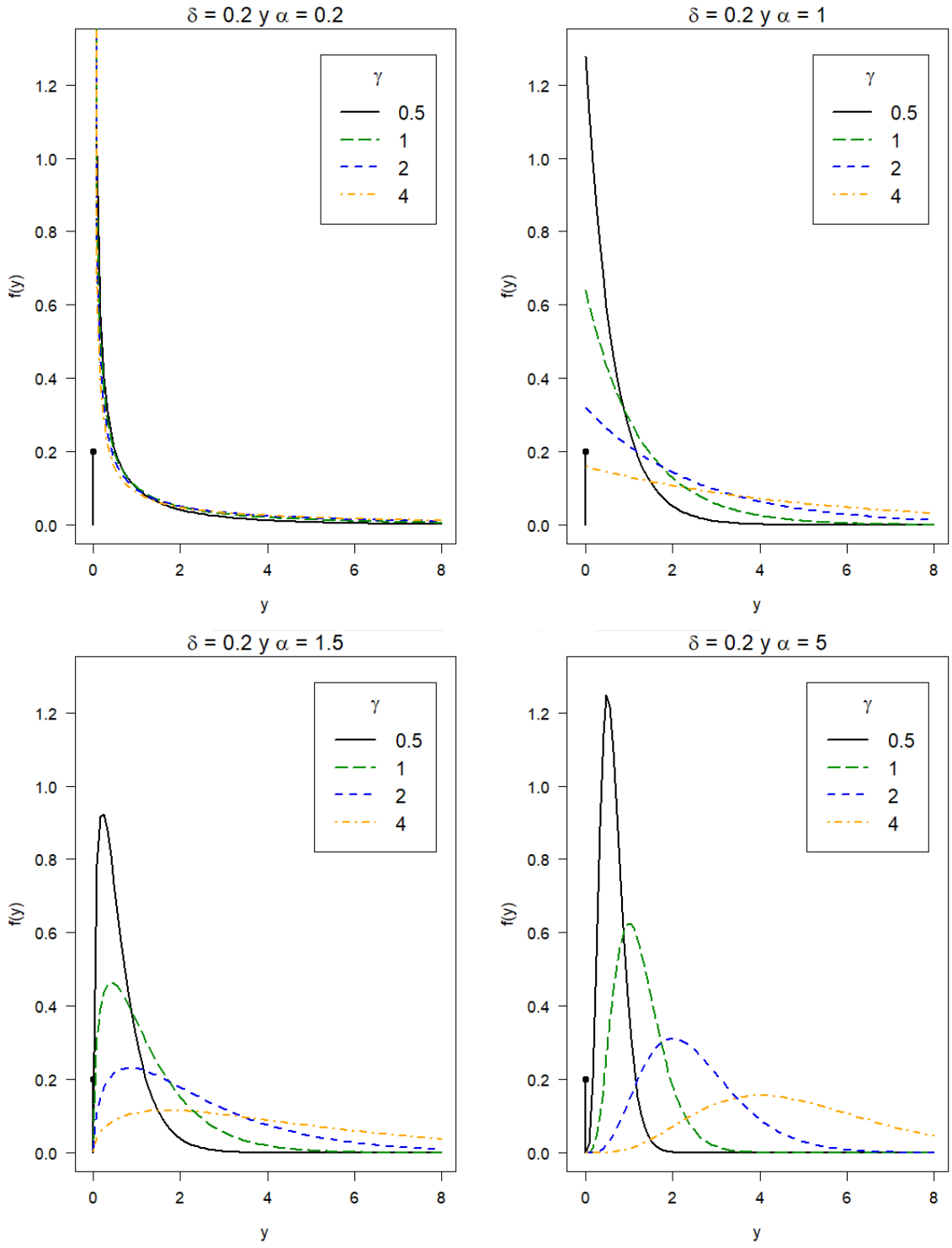


Figura 2.1: Función de masa de la distribución gamma cero-inflacionada bajo la parametrización de (2.15) para diferentes valores de los parámetros γ y α

2.3. La distribución gamma generalizada

Es una distribución de probabilidades de tres parámetros aplicable a variables continuas y no negativas, de carácter flexible dado que tiene como casos particulares a un número considerable de distribuciones, muchas de ellas conocidas y potencialmente útiles en el modelamiento de variables semicontinuas, tales como las distribuciones exponencial, log-normal, Weibull, gamma estándar, gamma inversa, entre otras.

Una de sus formulaciones iniciales es la propuesta por Stacy (1962) que consiste en incluir un segundo parámetro de forma, $\rho > 0$, en la distribución gamma estándar de dos parámetros mediante:

$$gg(y \mid \alpha, \beta, \rho) = \frac{1}{\Gamma(\alpha)} \left(\frac{1}{\beta}\right)^{\alpha\rho} \rho y^{\alpha\rho-1} \exp\left(-\left(\frac{y}{\beta}\right)^{\rho}\right), \qquad y \geq 0. \tag{2.18}$$

En Stacy y Mihram (1965) se propone extender el espacio paramétrico de ρ hacia valores negativos, siendo ahora la única restricción que $\rho \neq 0$. Ello se logra reemplazando en (2.18) el factor $\rho z^{\alpha\rho-1}$ por $|\rho| z^{\alpha\rho-1}$ y conservando los espacios paramétricos de $\alpha > 0$ y $\beta > 0$.

Una parametrización alternativa es presentada en Manning et al. (2005), quienes, para estimar un modelo de regresión por máxima verosimilitud, establecen que α, β y ρ sean expresados en función de nuevos parámetros λ, κ y σ de la siguiente manera:

$$\alpha = \frac{1}{|\kappa|^2} \tag{2.19}$$

$$\beta = \frac{\exp(\lambda)}{|\kappa|^{-2\,\text{Signo}(\kappa)\,\sigma/|\kappa|}} \tag{2.20}$$

$$\rho = \text{Signo}(\kappa)\frac{|\kappa|}{\sigma} \tag{2.21}$$

donde Signo(κ) es una función que toma el valor de 1 si $\kappa > 0$ y toma -1 si $\kappa < 0$.

Reemplazando estos nuevos parámetros en (2.18), como se ilustra en el Apéndice A.1, se puede decir que una variable aleatoria continua presenta distribución gamma generalizada de parámetros $-\infty < \lambda < \infty$, $\kappa \neq 0$ y $\sigma > 0$, si su función de densidad está dada por:

$$gg(y \mid \lambda, \kappa, \sigma) = \frac{|\kappa|^{-2/|\kappa|^2}}{\sigma y |\kappa|^{-1}\Gamma(1/|\kappa|^2)} \exp\left(\frac{\nu}{|\kappa|} - \frac{\exp(|\kappa|\nu)}{|\kappa|^2}\right), \qquad y \geq 0, \tag{2.22}$$

donde $\nu = \text{Signo}(\kappa)\left[\dfrac{\ln(y) - \lambda}{\sigma}\right]$, λ es un parámetro de localización, κ es un parámetro de forma y σ es un parámetro de escala.

El Cuadro 2.1 muestra cuáles son los valores que debe tomar los parámetros κ y σ para verificar algunos casos especiales de la gamma generalizada. Por ejemplo, la distribución gamma es un caso especial de la distribución gamma generalizada cuando $\kappa = \sigma$.

En esta tesis consideraremos como parte del modelo de distribución de la variable res-

Cuadro 2.1: Casos especiales de la distribución gamma generalizada

Distribución	κ	σ	λ
Exponencial	$\kappa = 1$	$\sigma = 1$	$-\infty < \lambda < \infty$
Gamma estándar	$\kappa = \sigma$	$\sigma > 0$	$-\infty < \lambda < \infty$
Log-normal	$\kappa \to 0$	$\sigma > 0$	$-\infty < \lambda < \infty$
Weibull	$\kappa = 1$	$\sigma > 0$	$-\infty < \lambda < \infty$
Gamma inversa	$\kappa = -\sigma$	$\sigma > 0$	$-\infty < \lambda < \infty$

puesta a la distribución gamma generalizada en términos de la parametrización propuesta por Manning et al. (2005), pero con un espacio paramétrico de κ restringido al caso donde toma valores positivos, $\kappa > 0$. Considerando ello, y luego de realizar una reexpresión que detallamos en el Apéndice A.1, decimos que una variable Y tiene distribución gamma generalizada de parámetros $-\infty < \lambda < \infty$, $\kappa > 0$ y $\sigma > 0$ si su función de densidad resulta dada por:

$$gg(y \mid \lambda, \kappa, \sigma) = \frac{\kappa/\sigma}{\Gamma(1/\kappa^2)} y^{\frac{1}{\kappa\sigma}-1} \exp\left(-\frac{\eta}{\kappa^2} y^{\frac{\kappa}{\sigma}} + \frac{1}{\kappa^2} \ln\left(\frac{\eta}{\kappa^2}\right)\right), \quad y \geq 0, \tag{2.23}$$

donde $\eta = \exp\left(-\frac{\kappa\lambda}{\sigma}\right)$.

La media y varianza de esta distribución están dadas respectivamente por:

$$E(Y) = \exp(\lambda)\kappa^{\frac{2\sigma}{\kappa}} \frac{\Gamma(1/\kappa^2 + \sigma/\kappa)}{\Gamma(1/\kappa^2)} \tag{2.24}$$

$$V(Y) = (\exp(\lambda)\kappa^{\frac{2\sigma}{\kappa}})^2 \left[\frac{\Gamma(1/\kappa^2 + 2\sigma/\kappa)}{\Gamma(1/\kappa^2)} - \left(\frac{\Gamma(1/\kappa^2 + \sigma/\kappa)}{\Gamma(1/\kappa^2)}\right)^2\right] \tag{2.25}$$

Los momentos de esta distribución están dados por:

$$E(Y^m) = (\exp(\lambda)\kappa^{\frac{2\sigma}{\kappa}})^m \frac{\Gamma(1/\kappa^2 + m\sigma/\kappa)}{\Gamma(1/\kappa^2)} \tag{2.26}$$

La forma de la distribución está enteramente dependiente de la combinación de valores que toma κ y σ. Utilizando los coeficientes de Fisher de asimetría y curtosis, observamos que los parámetros κ y σ tienen influencia sobre la asimetría y la curtosis de esta distribución; en cambio, el parámetro λ no tiene influencia, lo cual era esperable dado que es un parámetro de localización. Los coeficientes para la distribución $GG(\lambda, \kappa, \sigma)$ están dados respectivamente por:

$$A(Y) = \frac{\tau(3) - 3\tau(1)\tau(2) + 2\tau(1)^3}{(\tau(2) - \tau(1)^2)^{3/2}} \tag{2.27}$$

$$K(Y) = \frac{\tau(4) - 4\tau(1)\tau(3) + 6\tau(1)^2\tau(2) - 3\tau(1)^4}{(\tau(2) - \tau(1)^2)^{4/2}} \tag{2.28}$$

donde $\tau(m) = \dfrac{\Gamma(1/\kappa^2 + m\sigma/\kappa)}{\Gamma(1/\kappa^2)}$.

La Figura 2.2 muestra el cambio de los coeficientes de asimetría y curtosis, para distintos valores de κ y σ. El comportamiento de la curtosis es parecida a la de una función convexa respecto a κ: la curtosis es muy alta cuando κ toma valores bajos o cercanos a cero, luego alcanza un nivel mínimo y, finalmente, aumenta conforme aumenta el valor de κ. La asimetría positiva también es afectada por κ pero no de forma importante, en comparación del efecto que tiene sobre la curtosis. Por otro lado, el parámetro σ tiene influencia positiva y fuerte sobre la asimetría positiva y el carácter leptocúrtico de la distribución. Cuando σ toma un valor cercano a 0, las mediciones indican que la distribución es aproximadamente simétrica.

La Figura 2.3 muestra todos los valores posibles que estos coeficientes de asimetría y curtosis toman en el subespacio paramétrico definido en $\kappa = [0.3; 3]$ y $\sigma = [0.3; 3]$. Verificamos que para valores cercanos a los límites superiores de este subespacio, la curtosis y la asimetría toman valores sumamente altos, que, junto a lo visto en la Figura 2.2, es producto sobre todo del cambio del parámetro σ.

De forma similar a lo hecho en (2.5), la distribución gamma generalizada puede ser reparametrizada para que esté explícitamente en función de la media de la variable. Para ello, consideraremos un parámetro μ dado por:

$$\mu = \exp(\lambda)\kappa^{\frac{2\sigma}{\kappa}} \frac{\Gamma(1/\kappa^2 + \sigma/\kappa)}{\Gamma(1/\kappa^2)} \tag{2.29}$$

A partir de (2.29), obtenemos una expresión de λ en función de los demás parámetros, incluyendo el parámetro μ:

$$\lambda = \ln(\mu) - \frac{2\sigma}{k}\ln(k) + \ln\left[\Gamma\left(\frac{1}{k^2}\right)\right] - \ln\left[\Gamma\left(\frac{1}{k^2} + \frac{\sigma}{k}\right)\right] \tag{2.30}$$

Esta expresión es luego reemplazada en la función de densidad (2.23). La expresión resultante es la función de densidad de la gamma generalizada de parámetros $0 < \delta < 1$, $\mu > 0$, $\kappa > 0$ y $\sigma > 0$:

$$gg(y \mid \mu, \kappa, \sigma) = \frac{\kappa/\sigma}{\Gamma(1/\kappa^2)} z^{\frac{1}{\kappa\sigma}-1} \exp\left(-\frac{\eta}{\kappa^2} z^{\frac{\kappa}{\sigma}} + \frac{1}{\kappa^2}\ln\left(\frac{\eta}{\kappa^2}\right)\right), \quad y \geq 0 \tag{2.31}$$

donde

$$\eta = \exp\left(-\frac{\kappa\lambda}{\sigma}\right)$$

$$\lambda = \ln(\mu) - \frac{2\sigma}{k}\ln(k) + \ln\left[\Gamma\left(\frac{1}{k^2}\right)\right] - \ln\left[\Gamma\left(\frac{1}{k^2} + \frac{\sigma}{k}\right)\right]$$

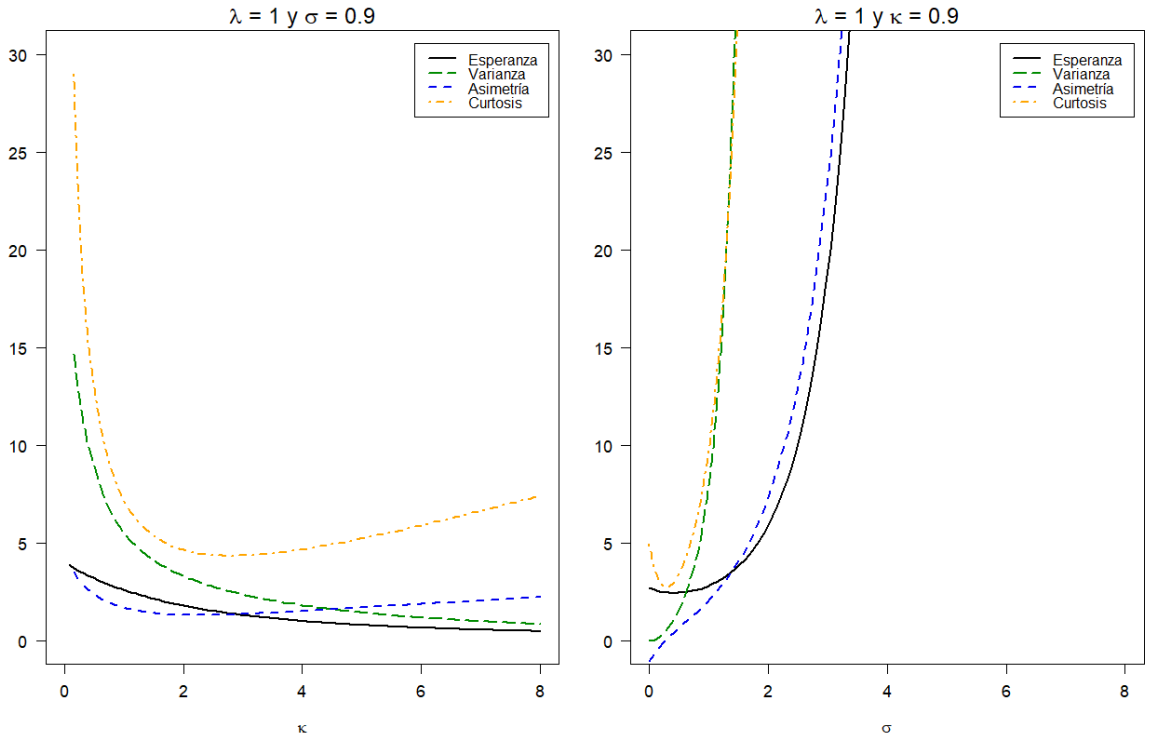


Figura 2.2: Esperanza, varianza, asimetría y curtosis de la distribución gamma generalizada en términos de (2.23) para diferentes valores de κ y σ.

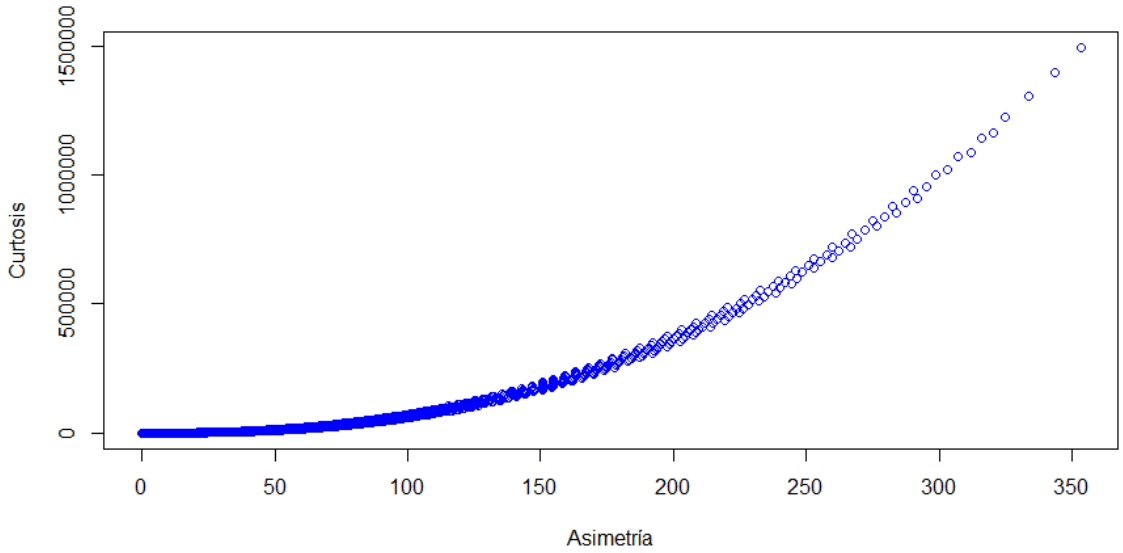


Figura 2.3: Coeficientes de Fisher de asimetría y curtosis de la distribución gamma generalizada en términos de (2.23) en el subespacio paramétrico $\kappa = [0.3; 3]$ y $\sigma = [0.3; 3]$. El lado izquierdo muestra todos los valores posibles de los coeficientes en aquel subespacio paramétrico. El lado derecho muestra un acercamiento.

2.4. La distribución gamma generalizada cero-inflacionada

Una variable aleatoria continua y no negativa Y tiene distribución gamma generalizada cero-inflacionada de parámetros $0 < \delta < 1$, $\mu > 0$, $\kappa > 0$ y $\sigma > 0$ si su función de masa de probabilidad está dada por:

$$f(y \mid \delta, \mu, \kappa, \sigma) = \begin{cases} \delta & , \text{ si } y = 0 \\ (1-\delta)\, gg(y \mid \mu, \kappa, \sigma) & , \text{ si } y > 0 \end{cases} \tag{2.32}$$

donde δ es la probabilidad que la variable tome el valor 0 y $gg(y \mid \delta, \mu, \kappa, \sigma)$ es la función densidad de la gamma generalizada en términos de (2.31). En adelante, si alguna variable aleatoria Y presenta distribución gamma generalizada en términos de (2.32), ello será denotado por $Y \sim GGCI(\delta, \mu, \kappa, \sigma)$.

La media y varianza de esta distribución cero-inflacionada están dadas por:

$$E(Y) = (1-\delta)\mu = (1-\delta)\exp(\lambda)\kappa^{\frac{2\sigma}{\kappa}} \frac{\Gamma(1/\kappa^2 + \sigma/\kappa)}{\Gamma(1/\kappa^2)} \tag{2.33}$$

$$V(Y) = (1-\delta)(\exp(\lambda)\kappa^{\frac{2\sigma}{\kappa}})^2 \left\{ \frac{\Gamma(1/\kappa^2 + 2\sigma/\kappa)}{\Gamma(1/\kappa^2)} - \left[\frac{\Gamma(1/\kappa^2 + \sigma/\kappa)}{\Gamma(1/\kappa^2)} \right]^2 (1-\delta) \right\} \tag{2.34}$$

donde λ toma la expresión de (2.31).

Estableceremos un nuevo parámetro γ definido como la media de la distribución cero-inflacionada, y que sustituirá al parámetro de localización, λ, de forma similar a lo efectuado en (2.13).

$$\gamma = (1-\delta)\,\mu = (1-\delta)\,\exp(\lambda)\kappa^{\frac{2\sigma}{\kappa}} \frac{\Gamma(1/\kappa^2 + \sigma/\kappa)}{\Gamma(1/\kappa^2)} \tag{2.35}$$

La función de densidad de la distribución gamma generalizada en términos de la nueva parametrización está dada por:

$$gg(y \mid \delta, \gamma, \kappa, \sigma) = \frac{\kappa/\sigma}{\Gamma(1/\kappa^2)} y^{\frac{1}{\kappa\sigma}-1} \exp\left(-\frac{\eta}{\kappa^2} z^{\frac{\kappa}{\sigma}} + \frac{1}{\kappa^2} \ln\left(\frac{\eta}{\kappa^2}\right) \right) \tag{2.36}$$

donde

$$\eta = \exp\left(-\frac{\kappa\lambda}{\sigma}\right)$$

$$\lambda = \ln(\gamma) - \ln(1-\delta) - \frac{2\sigma}{k}\ln(k) + \ln\left[\Gamma\left(\frac{1}{k^2}\right)\right] - \ln\left[\Gamma\left(\frac{1}{k^2} + \frac{\sigma}{k}\right)\right]$$

Entonces ahora una variable aleatoria Y tiene una distribución gamma generalizada cero-inflacionada en términos de la parametrización $0 < \delta < 1$, $\gamma > 0$, $\kappa > 0$ y $\sigma > 0$ si su función

de masa de probabilidad está dada por:

$$f(y \mid \delta, \gamma, \kappa, \sigma) = \begin{cases} \delta & , \text{ si } y = 0 \\ (1-\delta)\, gg(y \mid \delta, \gamma, \kappa, \sigma) & , \text{ si } y > 0 \end{cases} \tag{2.37}$$

donde $gg(y \mid \delta, \gamma, \kappa, \sigma)$ es la función de densidad (2.36). En adelante denotaremos a una variable Y que presenta distribución gamma generalizada cero-inflacionada en términos de (2.37) por $Y \sim GGCI(\delta, \gamma, \kappa, \sigma)$.

La forma de la distribución $GGCI(\delta, \gamma, \kappa, \sigma)$ está dependiente de la combinación de valores que toma los parámetros δ, κ y σ. Utilizando los coeficientes de Fisher de asimetría y curtosis, observamos que estos parámetros tienen influencia sobre la asimetría y la curtosis de esta distribución. Los coeficientes para la distribución $GGCI(\delta, \gamma, \kappa, \sigma)$ están dados respectivamente por:

$$A(Y) = \frac{\tau(\delta,3) - 3\tau(\delta,1)\tau(\delta,2) + 2\tau(\delta,1)^3}{(\tau(\delta,2) - \tau(\delta,1)^2)^{3/2}} \tag{2.38}$$

$$K(Y) = \frac{\tau(\delta,4) - 4\tau(\delta,1)\tau(\delta,3) + 6\tau(\delta,1)^2\tau(\delta,2) - 3\tau(\delta,1)^4}{(\tau(\delta,2) - \tau(\delta,1)^2)^{4/2}} \tag{2.39}$$

donde $\tau(\delta, m) = (1-\delta)\dfrac{\Gamma(1/\kappa^2 + m\sigma/\kappa)}{\Gamma(1/\kappa^2)}$.

La Figura 2.4 muestra el cambio de los coeficientes de Fisher de asimetría y de curtosis para distintos valores de κ y σ. En el caso de δ y γ, las mediciones tienen comportamiento similar que el caso gamma generalizada cero-inflacionada. Las Figuras 2.5 y 2.6 muestran la forma de la función de densidad para distintos valores de sus parámetros. Observamos que la forma de la distribución se aproxima a una simétrica cuando κ y σ toman valores cercanos a cero. Cuando $\sigma > 1$, la función inicia en infinito y cae rápidamente en forma de L, independientemente del valor que tome κ. Cuando $\kappa < 1$ y $\sigma < 1$, la función inicia en un punto cercano a cero.

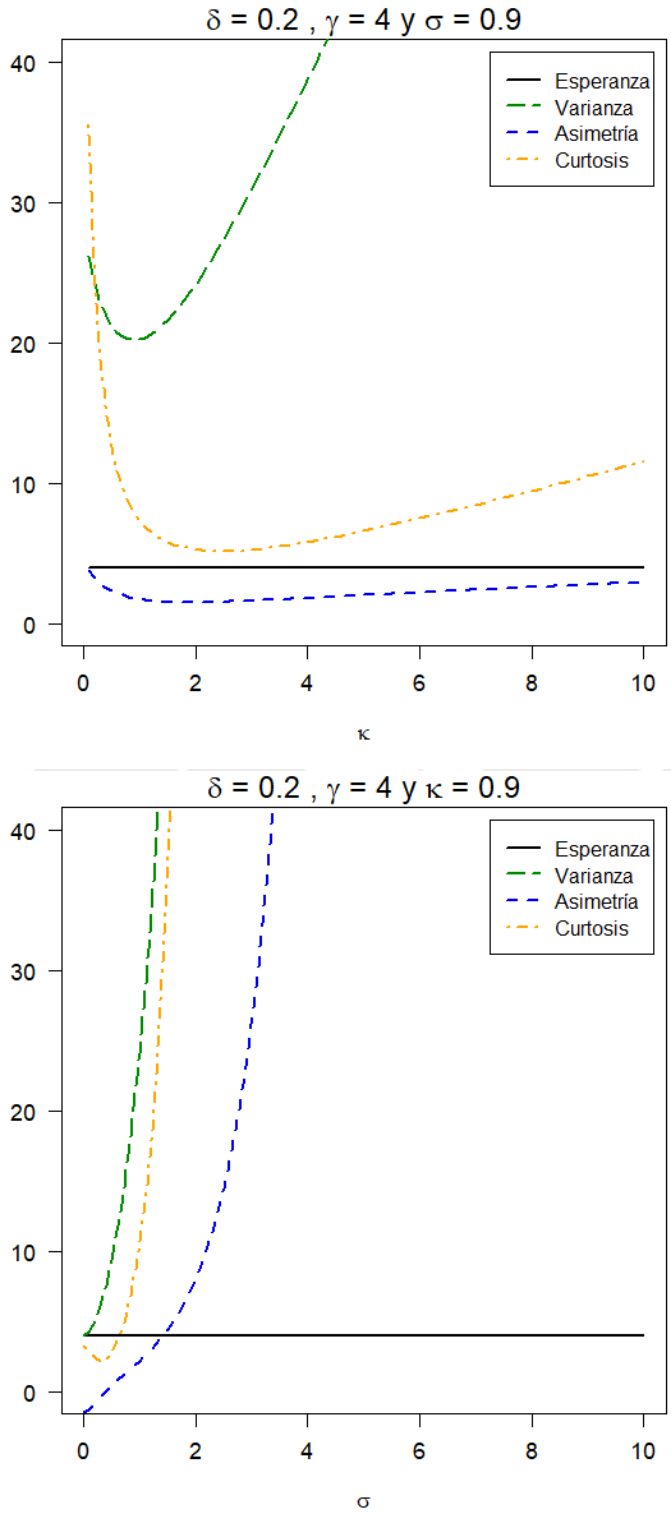


Figura 2.4: Esperanza, varianza, asimetría y curtosis de la distribución GGCI(δ, γ, κ, σ) para diferentes valores de κ y σ. En el lado izquierdo, κ varía y el resto de parámetros es constante. En el lado derecho, σ varía y el resto es constante

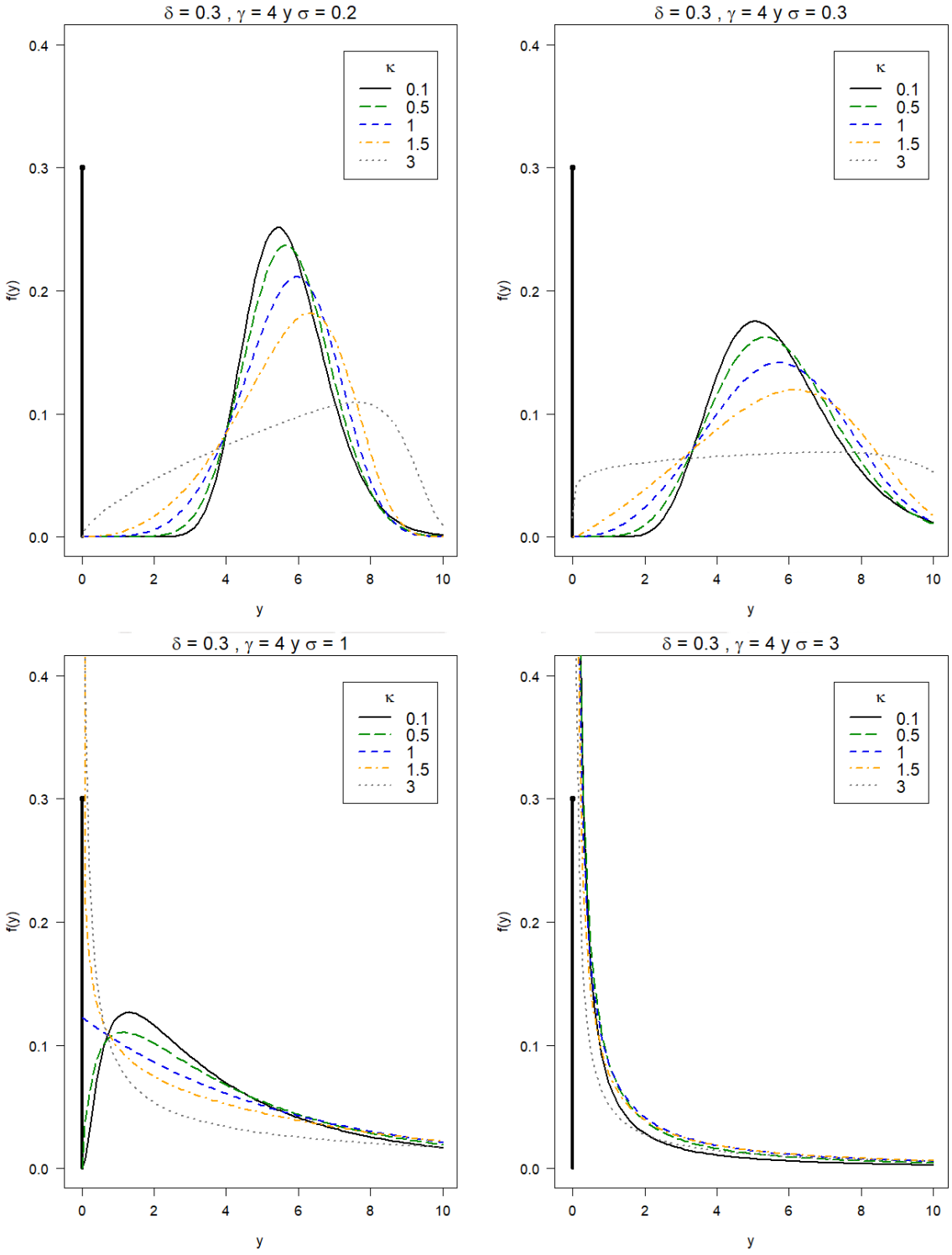


Figura 2.5: Función de masa de la distribución GGCI(δ, γ, κ, σ) para diferentes valores de κ y σ

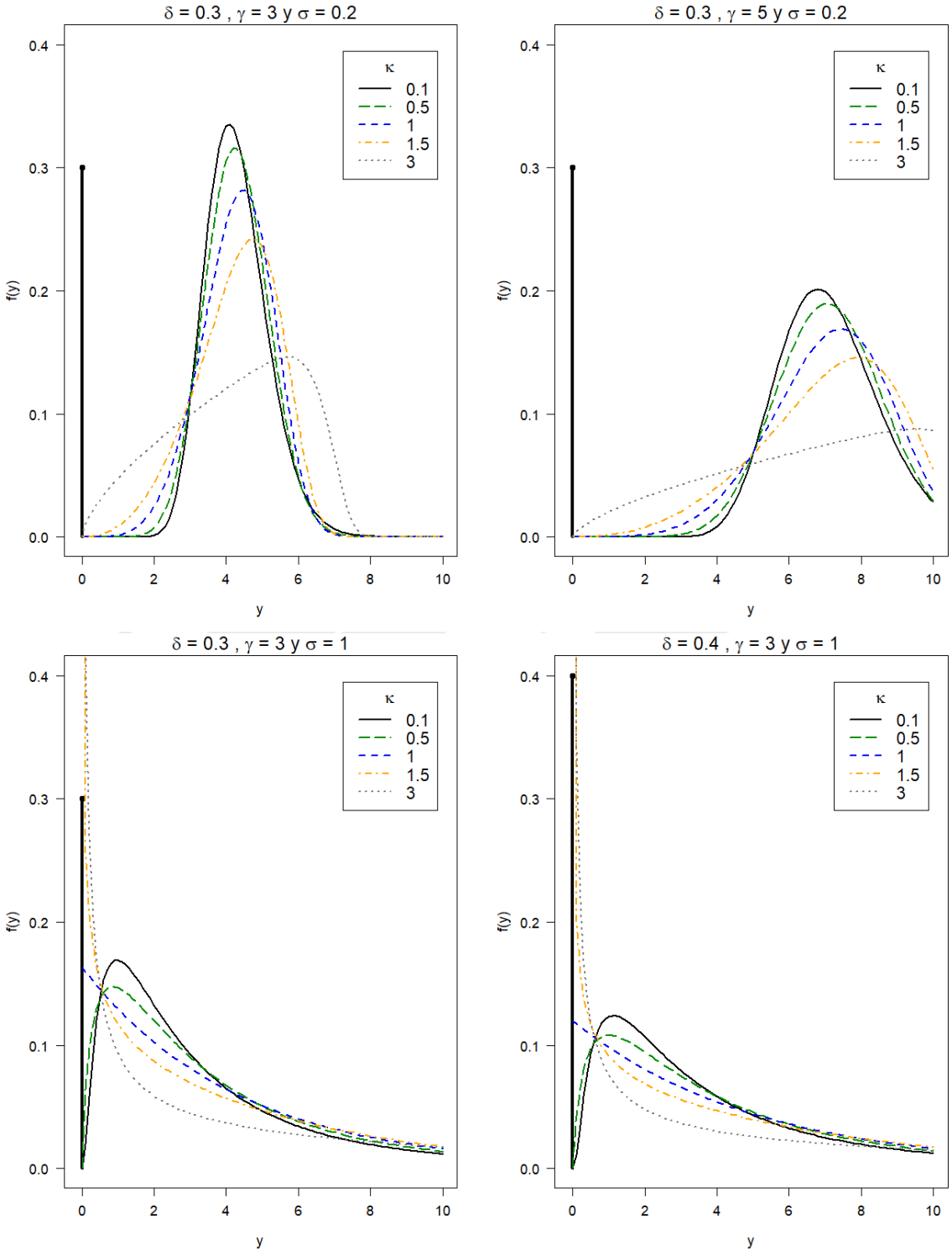


Figura 2.6: Función de masa de la distribución GGCI(δ, γ, κ, σ) para diferentes valores de κ, δ y γ

Capítulo 3

Modelos de regresión para respuesta semicontinua

En este capítulo describiremos las principales características de dos modelos de regresión utilizados para explicar una variable semicontinua: el modelo de regresión de dos partes y el modelo de regresión cero-inflacionada de una parte.

3.1. El modelo de regresión de dos partes

En el análisis de regresión, el modelo de dos partes (MDP) es el tradicionalmente usado para explicar una variable semicontinua. Fue presentado por Duan et al. (1983) y plantea modelar la respuesta semicontinua en dos partes. La primera parte consiste en una regresión binaria con la finalidad de explicar la presencia o no de valores cero de la respuesta. Usando solamente los valores positivos, la segunda parte es una regresión para explicar el nivel de la respuesta positiva con, por ejemplo, una regresión gamma o log-normal.

En el MDP la respuesta presenta distribución del tipo cero-inflacionada y separadamente se estima los efectos de covariables sobre la probabilidad de que la respuesta tome el valor cero y los efectos de covariables sobre la media de la respuesta condicionada a valores positivos. Sea $Y_1, Y_2, ..., Y_n$ un conjunto de n variables aleatorias independientes y con idéntica distribución y sea y_i el valor que toma la respuesta Y_i del i-ésimo sujeto, donde $i = 1, 2, ..., n$. La función de masa de probabilidad de cada Y_i está dada por:

$$f(y \,|\, \delta, \mu, \boldsymbol{\psi}) = \begin{cases} \delta & , \text{ si } y = 0 \\ (1-\delta)\, h(y \,|\, \mu, \boldsymbol{\psi}) & , \text{ si } y > 0 \end{cases} \tag{3.1}$$

donde $\delta = P(Y = 0)$, $\mu = E(Y|Y > 0)$ y $h(y \,|\, \mu, \boldsymbol{\psi})$ es una función de densidad de alguna distribución condicionada a valores positivos cuyos parámetros son μ y otros contenidos en el vector de parámetros $\boldsymbol{\psi}$.

Además, en el MDP se plantea las siguientes ecuaciones de regresión para δ_i y μ_i:

$$e_1(\delta_i) = \widetilde{\boldsymbol{x}_i^t \boldsymbol{\omega}} \tag{3.2}$$

$$e_2(\mu_i) = \boldsymbol{x}_i^t \boldsymbol{\beta} \tag{3.3}$$

donde $\boldsymbol{\omega} = (\omega_0, \omega_1, \omega_2, ..., \omega_{k_1})^t$ y $\boldsymbol{\beta} = (\beta_0, \beta_1, \beta_2, ..., \beta_{k_2})^t$ son vectores columna de los parámetros de regresión, $\widetilde{\boldsymbol{x}}_{\boldsymbol{i}}^{\boldsymbol{t}} = (\widetilde{x}_{i1}, \widetilde{x}_{i2}, ..., \widetilde{x}_{ik_1})$ y $\boldsymbol{x}_{\boldsymbol{i}}^{\boldsymbol{t}} = (x_{i1}, x_{i2}, ..., x_{ik_2})$ son conjuntos de covariables para el sujeto i-ésimo que están relacionadas con sus parámetros δ_i y μ_i, respectivamente; y e_1 y e_2 son funciones de enlace apropiadas.

Es importante mencionar que por medio de la mixtura de distribuciones se logra que distribuciones aplicables a valores positivos puedan ser útiles para modelar respuestas semicontinuas. Recordar que distribuciones como, por ejemplo, la gamma, la log-normal o la Weibull, tienen dominios limitados al intervalo abierto $(0, \infty)$, que no incluye al valor cero. Al considerarlas en la mixtura se aprovecha la ventaja que tienen de modelar asimetría positiva y cola pesada. Asimismo, la mixtura permite que la parametrización se plantee de tal forma que los valores positivos sean explicados por covariables que no son necesariamente las mismas que explican los valores cero de la respuesta.

3.2. El modelo de regresión cero-inflacionada a la media

El interés del análisis de regresión también puede centrarse en estimar directamente e interpretar los efectos de covariables sobre la media total de la respuesta, $\gamma = E(Y) = (1-\delta)\mu$, es decir, la media marginal que incluye los valores cero y los valores positivos; en lugar de la media condicionada a valores positivos, $\mu = E(Y|Y > 0)$. Una forma de abordar este interés es cambiar la parametrización del modelo para que la función de densidad sea expresada en términos γ, y luego plantear su ecuación de regresión. Llamaremos a esta propuesta como modelo de regresión cero-inflacionada a la media (MCIM).

Cabe precisar que el modelo con esta reparametrización deja de ser de dos partes, porque el nombre se debía a la intención o posibilidad de estimar de forma separada, en dos partes, a $\gamma = (1 - \delta)\mu$. Ahora es adecuado nombrarlo como regresión de una parte, porque éste se estimará en un solo proceso de optimización.

En Bayes y Valdivieso (2016) se propone este tipo de reparametrización para una distribución beta inflacionada con variables fraccionadas acotadas (que toma valores en el intervalo cerrado [0,1]) sin que el espacio paramétrico de γ se vea restringido. En Smith et al. (2014) se muestra esta reparametrización para distribuciones log-normal cero-inflacionada y log-skew-normal cero-inflacionada con variables semicontinuas. En ambos estudios el modelo conserva su capacidad de obtener estimaciones de efectos de covariables sobre δ.

En concreto, en el MCIM se establece que la función de masa de probabilidad de Y_i esté dada ahora por:

$$f(y \,|\, \delta, \gamma, \boldsymbol{\psi}) = \begin{cases} \delta & , \text{ si } y = 0 \\ (1-\delta)\; h(y \,|\, \gamma, \boldsymbol{\psi}) & , \text{ si } y > 0 \end{cases} \tag{3.4}$$

donde $\delta = P(Y = 0)$, $\gamma = E(Y)$ y $h(y \,|\, \gamma, \boldsymbol{\psi})$ es una función de densidad de alguna distribución condicionada a valores positivos cuyos parámetros son γ y otros contenidos en el vector $\boldsymbol{\psi}$.

En el MCIM las ecuaciones de regresión para δ_i y γ_i son planteadas ahora como:

$$e_1(\delta_i) = \widetilde{\boldsymbol{x}_i^t}\boldsymbol{\omega} \tag{3.5}$$

$$e_2(\gamma_i) = \boldsymbol{x}_i^t\boldsymbol{\beta} \tag{3.6}$$

Para los fines de la tesis, consideraremos que los parámetros δ_i y γ_i serán explicados por predictores lineales usando funciones de enlace estrictamente monótonas y con segunda derivada continua. Más aún, con el motivo de facilitar la interpretación de los coeficientes de regresión, elegiremos como función de enlace g_1 a la logística y como g_2 a la logarítmica:

$$e_1(\delta_i) = \ln\left(\frac{\delta_i}{1-\delta_i}\right) = \widetilde{\boldsymbol{x}_i^t}\boldsymbol{\omega} \tag{3.7}$$

$$e_2(\gamma_i) = \ln(\gamma_i) = \boldsymbol{x}_i^t\boldsymbol{\beta} \tag{3.8}$$

Además, especificaremos en esta tesis que $Y_i \sim GGCI(\delta_i, \gamma_i, \kappa, \sigma)$, en los términos descritos en (2.37), lo cual indica que el componente continuo positivo de esta variable tiene distribución gamma generalizada en términos de (2.36). Más explícitamente, la función de masa de probabilidad de la variable respuesta Y_i estará dado por:

$$f(y \mid \delta_i, \gamma_i, \kappa, \sigma) = \begin{cases} \delta_i & , \text{ si } y = 0 \\ (1-\delta_i)\,\dfrac{\kappa/\sigma}{\Gamma(1/\kappa^2)} y^{\frac{1}{\kappa\sigma}-1} \exp\left(-\dfrac{\eta_i}{\kappa^2}y^{\frac{\kappa}{\sigma}} + \dfrac{1}{\kappa^2}\ln\left(\dfrac{\eta_i}{\kappa^2}\right)\right) & , \text{ si } y > 0 \end{cases} \tag{3.9}$$

donde

$$\eta_i = \exp\left(-\frac{\kappa\lambda_i}{\sigma}\right)$$

$$\lambda_i = \ln(\gamma_i) - \ln(1-\delta_i) - \frac{2\sigma}{\kappa}\ln(\kappa) + \ln\left[\Gamma\left(\frac{1}{\kappa^2}\right)\right] - \ln\left[\Gamma\left(\frac{1}{\kappa^2} + \frac{\sigma}{\kappa}\right)\right]$$

La estimación del vector de parámetros de regresión $\boldsymbol{\theta} = [\boldsymbol{\omega}, \boldsymbol{\beta}, \kappa, \sigma]^t$, se realizará por el método de máxima verosimilitud. Sin incurrir en pérdida de generalidad, asumimos que las n observaciones de la variable respuesta estarán ordenadas de tal manera que las primeras n_0 observaciones toman el valor cero y las restantes $n - n_0$ toman un valor positivo. Por tanto, la función de verosimilitud estará dada por:

$$\mathbb{L}(\boldsymbol{\theta}) = \prod_{i=1}^{n} f(y_i \mid \delta_i, \gamma_i, \kappa, \sigma) \tag{3.10}$$

$$= \prod_{i=1}^{n_0} \delta_i \prod_{i=n_0+1}^{n} (1-\delta_i)\, gg(y_i \mid \delta_i, \gamma_i, \kappa, \sigma) \tag{3.11}$$

$$= \prod_{i=1}^{n_0} \delta_i \prod_{i=n_0+1}^{n} (1-\delta_i)\,\frac{\kappa/\sigma}{\Gamma(1/\kappa^2)}\, y_i^{\frac{1}{\kappa\sigma}-1}\, \exp\left(-\frac{\eta_i}{\kappa^2}y_i^{\frac{\kappa}{\sigma}} + \frac{1}{\kappa^2}\ln\left(\frac{\eta_i}{\kappa^2}\right)\right) \tag{3.12}$$

donde

$$\delta_i = \frac{\exp(\widetilde{\boldsymbol{x}}_{\boldsymbol{i}}^{\boldsymbol{t}}\boldsymbol{\omega})}{1+\exp(\widetilde{\boldsymbol{x}}_{\boldsymbol{i}}^{\boldsymbol{t}}\boldsymbol{\omega})}$$
$$\gamma_i = \exp(\boldsymbol{x}_{\boldsymbol{i}}^{\boldsymbol{t}}\boldsymbol{\beta})$$
$$\eta_i = \exp\left(-\frac{\kappa\lambda_i}{\sigma}\right)$$
$$\lambda_i = \ln(\gamma_i) - \ln(1-\delta_i) - \frac{2\sigma}{\kappa}\ln(\kappa) + \ln\left[\Gamma\left(\frac{1}{\kappa^2}\right)\right] - \ln\left[\Gamma\left(\frac{1}{\kappa^2}+\frac{\sigma}{\kappa}\right)\right]$$

Por tanto, la función de log-verosimilitud del MCIM-GG está dada por:

$$\mathbb{K}(\boldsymbol{\theta}) = \sum_{i=1}^{n_0} \ln\,\delta_i + \sum_{i=n_0+1}^{n} \ln(1-\delta_i) + \sum_{i=n_0+1}^{n} \ln[gg(y_i \,|\, \delta_i, \gamma_i, \kappa, \sigma)] \tag{3.13}$$

donde

$$\begin{aligned}\ln[gg(y_i \,|\, \delta_i, \gamma_i, \kappa, \sigma)] =& \ln(\kappa) - \ln(\sigma) + \left(\frac{1}{\sigma k} - 1\right)\ln(y_i) - \ln\left[\Gamma\left(\frac{1}{\kappa^2}\right)\right] - \frac{\eta_i}{\kappa^2} y_i^{\frac{\kappa}{\sigma}} \\ &+ \frac{1}{\kappa^2}\ln(\eta_i) - \frac{1}{\kappa^2}\ln(k^2)\end{aligned}$$

Un caso particular de este modelo es el MCIM-G, donde el componente continuo Z_i presenta una distribución gamma estándar cuya parametrización está en términos de (2.15). En este caso, el vector de parámetros a estimar es $\boldsymbol{\theta} = (\boldsymbol{\omega}, \boldsymbol{\beta}, \alpha)^t$ y la función de verosimilitud está dada por:

$$\mathbb{L}(\boldsymbol{\theta}) = \prod_{i=1}^{n_0} \delta_i \prod_{i=n_0+1}^{n} (1-\delta_i)\,\frac{1}{\Gamma(\alpha)}\left[\frac{(1-\delta_i)\alpha}{\gamma_i}\right]^{\alpha} y_i^{\alpha-1}\,\exp\left(-y_i\frac{(1-\delta_i)\alpha}{\gamma_i}\right) \tag{3.14}$$

donde $\delta_i = \dfrac{\exp(\widetilde{\boldsymbol{x}}_{\boldsymbol{i}}^{\boldsymbol{t}}\boldsymbol{\omega})}{1+\exp(\widetilde{\boldsymbol{x}}_{\boldsymbol{i}}^{\boldsymbol{t}}\boldsymbol{\omega})}$ y $\gamma_i = \exp(\boldsymbol{x}_{\boldsymbol{i}}^{\boldsymbol{t}}\boldsymbol{\beta})$.

Por tanto, la función de log-verosimilitud del MCIM-G está dada por:

$$\mathbb{K}(\boldsymbol{\theta}) = \sum_{i=1}^{n_0} \ln\,\delta_i + \sum_{i=n_0+1}^{n} \ln(1-\delta_i) + \sum_{i=n_0+1}^{n} \ln[g(y_i \,|\, \delta_i, \gamma_i, \alpha)] \tag{3.15}$$

donde

$$\begin{aligned}\ln[g(y_i \,|\, \delta_i, \gamma_i, \alpha)] =& -\ln\left[\Gamma(\alpha)\right] - \alpha\ln\left(\frac{\mu_i}{\alpha}\right) + (\alpha-1)\ln(y_i) - y_i\left(\frac{\alpha}{\mu_i}\right) \\ \mu_i =& \frac{\gamma_i}{1-\delta_i}\,.\end{aligned}$$

La optimización de la log-verosimilitud del MCIM-GG se realizará utilizando la rutina "fmincon" del software Matlab (The MathWorks, Inc., 2018), el cual usará el vector gradiente

y la matriz hessiana analíticos. En el Apéndice A.2 mostramos el procedimiento para obtener las primeras y segundas derivadas de la función de log-verosimilitud del MCIM-GG. Éstas serán relevantes no solamente para obtener las estimaciones de máxima verosimilitud, sino también para los errores estándar de estimación a través del cálculo de la matriz de información de Fisher observada evaluada en las correspondientes estimaciones.

Para la optimización es necesario determinar estimaciones iniciales $\widehat{\boldsymbol{\theta}}_{\mathbf{0}} = [\widehat{\boldsymbol{\omega}}_{\mathbf{0}}, \widehat{\boldsymbol{\beta}}_{\mathbf{0}}, \widehat{\kappa_0}, \widehat{\sigma_0}]^t$ del vector de parámetros $\boldsymbol{\theta}$. Proponemos considerar como $\widehat{\boldsymbol{\omega}}_{\mathbf{0}}$ a las estimaciones de los coeficientes de una regresión logística para las n observaciones de Y expresadas en valores 0 y 1, donde las observaciones positivas de Y tomarán el valor 1. Por otro lado, consideramos como $\widehat{\boldsymbol{\beta}}_{\mathbf{0}}$ a las estimaciones de parámetros de una regresión gamma entre los $n-n_0$ valores positivos de Y y las covariables correspondientes. Finalmente, consideraremos que $\widehat{\kappa}_0 = \widehat{\sigma}_0 = \sqrt{1/\widehat{\alpha}_0}$, donde $\widehat{\alpha}_0$ es la estimación del parámetro de dispersión de la regresión gamma.

La optimización de la log-verosimilitud puede lograrse también con la rutina "proc nlmixed" del software SAS (SAS Institute Inc., 2018), dado que su tarea general es la estimación de modelos mixtos no lineales. En la aplicación usaremos esta rutina para la estimación del MDP-GG.

3.3. Selección del modelo

Los criterios utilizados en esta tesis para seleccionar el mejor modelo serán el Criterio de Información de Akaike (AIC), el Criterio de Información de Akaike corregido (AICc) y el Criterio de Información Bayesiano (BIC). Estos criterios tienen como característica común el premiar el buen ajuste del modelo a los datos (medido por el valor de la función de verosimilitud de los datos observados) y penalizar la complejidad del mismo (medido por el número de los parámetros a estimar). Entre los modelos candidatos para un conjunto dado de datos, el mejor modelo será quien registre el menor valor del criterio de información.

El AIC es definido por:

$$\text{AIC} = 2p - 2\ln\left(\mathbb{L}(\widehat{\boldsymbol{\theta}}_{\boldsymbol{MV}})\right) , \tag{3.16}$$

donde p es el número de parámetros estimados del modelo y $\mathbb{L}$ es la función de verosimilitud del modelo y $\widehat{\boldsymbol{\theta}}_{\boldsymbol{MV}}$ es el vector de parámetros estimados por el método de máxima verosimilitud.

El AICc es el AIC con una corrección para tamaños de muestra finitos:

$$\text{AICc} = \text{AIC} + \frac{2p(p+1)}{n-p-1} , \tag{3.17}$$

donde n es el número de observaciones de la muestra.

Por último, el BIC considera como parte del peso de penalización al número de observaciones de la muestra, siendo definido por:

$$\text{BIC} = \ln(n)p - 2\ln\left(\mathbb{L}(\widehat{\boldsymbol{\theta}}_{\boldsymbol{MV}})\right) . \tag{3.18}$$

Capítulo 4

Estudio de simulación

El objetivo de esta simulación será evaluar el desempeño de las estimaciones de máxima verosimilitud del MCIM-GG en diferentes escenarios de simulación.

4.1. Descripción

Primero, describimos las condiciones que se cumplirán para todos los escenarios de simulación. Respecto a las covariables, consideramos que $\delta = P(Y = 0)$ y $\gamma = E(Y)$ están relacionados con un único conjunto de cuatro covariables que serán simuladas de las siguientes distribuciones:

$$X_1 \sim U(a = 0, b = 1)$$
$$X_2 \sim N(\mu = 3, \sigma = 1)$$
$$X_3 \sim Bernoulli(p = 0.50)$$

La relación de éstas covariables con la media total de la respuesta, γ_i, estará dada por la siguiente especificación:

$$\ln(\gamma) = -1 - 2.9X_1 + 1.1X_2 - 1.1X_3$$

Seguidamente, usándose las especificaciones anteriores, simularemos datos de la la variable respuesta con distribución:

$$Y \sim GGCI(\delta, \gamma, \kappa = 0.50, \sigma = 0.70)$$

Esta distribución presenta asimetría positiva y una forma leptocúrtica más pronunciadas respecto a una distribución $Y \sim GGCI(\delta, \gamma, \kappa = 0.5, \sigma = 0.5)$, que equivale a una gamma estándar cero-inflacionada de parámetro $\alpha = 1/\sqrt{\kappa} = 1/\sqrt{0.5}$.

En tercer lugar, consideramos que las simulaciones presenten diferentes proporciones de valores ceros de la respuesta. Las bases de datos tendrán aproximadamente 10 %, 20 % y 40 %

de valores ceros respecto al total de valores observados de la respuesta. Esto se logrará si se establece que la relación entre las covariables y la probabilidad de que Y sea cero esté dada por las diferentes especificaciones:

$$\begin{aligned}\text{logit}\,(\delta) &= -1 + 3.1\,X_1 - 1.5\,X_2 + 1.3\,X_3\\ \text{logit}\,(\delta) &= -2 - 3.1\,X_1 + 0.9\,X_2 - 1.8\,X_3\\ \text{logit}\,(\delta) &= -3 - 1.2\,X_1 + 1.5\,X_2 - 2.9\,X_3\end{aligned}$$

que generan aproximadamente porcentajes de 10 %, 20 % y 40 %, respectivamente.

En cuarto lugar, para cada una de estos tres escenarios, consideramos que las simulaciones sean de diferentes tamaños de muestra: 200, 500, 1,000 y 3,000. Entonces, en forma de resumen, conformaremos 1,000 simulaciones o bases de datos en cada uno de los 12 escenarios que combinan los porcentajes de valores ceros de la respuesta y los tamaños de muestra.

Finalmente, ajustamos o estimamos los parámetros de regresión del MCIM-GG sobre el conjunto de valores de la respuesta y covariables de cada base de datos simulada. Culminado ello, verificamos el desempeño del modelo en recuperar los parámetros preestablecidos, en términos de sesgo, sesgo relativo, raíz del error cuadrático medio (RECM) y cobertura al 95 % del intervalo de confianza, definidos en cada escenario como:

$$\text{Sesgo}\left(\widehat{\theta}_j\right) = \frac{\sum_{i=1}^{m}\widehat{\theta}_j^{(i)}}{m} - \theta_j \tag{4.1}$$

$$\text{Sesgo\_Relativo}\left(\widehat{\theta}_j\right) = \frac{\text{Sesgo}\left(\widehat{\theta}_j\right)}{\theta_j}\,100\,\% \tag{4.2}$$

$$\text{RECM}\left(\widehat{\theta}_j\right) = \sqrt{\frac{\sum_{i=1}^{m}\left(\widehat{\theta}_j^{(i)} - \theta_j\right)^2}{m}} \tag{4.3}$$

$$\text{Cobertura}\left(\widehat{\theta}_j\right) = \frac{\sum_{i=1}^{m}\mathbb{I}\left[\theta_j \in \text{IC}\left(\widehat{\theta}_j^{(i)}\right)\text{ de }95\,\%\right]}{m}\,100\,\% \tag{4.4}$$

donde $\widehat{\theta}_j^{(i)}$ es la estimación del parámetro θ_j en la i-ésima simulación, m es la cantidad total de simulaciones, $\text{IC}\left(\widehat{\theta}_j^{(i)}\right)$ es un intervalo de confianza para θ_j en base a la estimación de máxima verosimilitud $\widehat{\theta}_j^{(i)}$, y $\mathbb{I}$ es una función indicadora que, para este caso, toma el valor de 1 si el $\text{IC}\left(\widehat{\theta}_j^{(i)}\right)$ al 95 % contiene a θ_j y 0 en caso contrario.

La maximización de la función log-verosimilitud del MCIM-GG es obtenida mediante la rutina "fmincon" del software Matlab, el cual tendrá como insumos al vector gradiente y a la matriz hessiana analíticos detallados en el Apéndice A.2.

4.2. Resultados

Los Cuadros 4.1, 4.2 y 4.3 muestran los resultados del estudio de simulación para los porcentaje de valores cero 10 %, 20 % y 40 %, respectivamente. Cada cuadro está dividido por los tamaño de muestra y muestran el sesgo, el sesgo porcentual, la RECM y la cobertura al 95 % de las 1,000 estimaciones obtenidas del MCIM-GG de cada escenario.

En relación al sesgo relativo, los estimadores de los parámetros de regresión de $P(Y = 0)$, estos son $\omega_0, \omega_1, \omega_2$ y ω_3, presentan sesgo relativos más altos, mayores que 6.5 %, en el escenario donde el tamaño de muestra es bajo, $n = 200$ y la proporción de ceros es bajo, aproximadamente 10 %. Esto era esperable dada la poca cantidad de observaciones cero que en tal escenario (aproximadamente 20 observaciones son cero) se le proporciona a la regresión de función logística. Luego, conforme aumenta el tamaño de muestra, estos sesgos relativos comienzan a disminuir. En el caso de los parámetros de regresión de $E(Y)$, estos son β_0, β_1, β_2 y β_3, sus sesgos relativos se mantuvieron en niveles bajos, menores que 1 %, en todos los escenarios de tamaños de muestra y de proporciones de ceros.

Observamos tendencias ligeramente distintas de las RECM de entre los parámetros de regresión de $P(Y = 0)$ y los de $E(Y)$ conforme cambia el porcentaje de valores cero de la respuesta. La RECM de los parámetros de $P(Y = 0)$ tiende a disminuir a medida que los porcentajes de ceros es mayor; en cambio, la RECM de los parámetros de $E(Y)$ tiende a aumentar. Por otro lado, observamos que la RECM de prácticamente todos los estimadores tienden a disminuir conforme aumenta el tamaño de muestra.

La intervalos de confianza de los parámetros de regresión contiene al verdadero valor del parámetro en un porcentaje cercano al 95 % de las simulaciones en cada escenario. Esta cercanía prácticamente sucede en todos los escenarios, independientemente del tamaño de muestra o del porcentaje de valores cero. Aunque, si separamos a los parámetros con cobertura menor que 94 %, estos se encuentran en mayor parte en el tamaño de muestra más bajo.

Cuadro 4.1: Resultados de simulación GGCI donde porcentaje de ceros 10 %

Tamaño de muestra	Parámetro	Valor verdadero	Sesgo	Sesgo relativo (%)	RECM	Cobertura al 95 % (%)
200	ω_0	-1.0	-0.0801	8.0	0.9133	97.0
	ω_1	3.1	0.2561	8.3	1.0826	96.5
	ω_2	-1.5	-0.0959	6.4	0.3341	96.0
	ω_3	1.3	0.0854	6.6	0.5994	93.7
	β_0	-1.0	-0.0070	0.7	0.2174	94.4
	β_1	-2.9	0.0021	-0.1	0.1868	95.3
	β_2	1.1	-0.0007	-0.1	0.0607	95.4
	β_3	-1.1	-0.0029	0.3	0.1085	95.3
	κ	0.5	0.0114	2.3	0.1980	95.7
	σ	0.7	-0.0175	-2.5	0.0451	93.2
500	ω_0	-1.0	-0.0390	3.9	0.6002	95.6
	ω_1	3.1	0.0642	2.1	0.6445	93.4
	ω_2	-1.5	-0.0233	1.6	0.1913	95.0
	ω_3	1.3	0.0377	2.9	0.3305	94.6
	β_0	-1.0	-0.0041	0.4	0.1376	94.4
	β_1	-2.9	0.0023	-0.1	0.1308	92.1
	β_2	1.1	0.0005	0.0	0.0388	94.3
	β_3	-1.1	-0.0028	0.3	0.0736	94.8
	κ	0.5	0.0035	0.7	0.1197	94.8
	σ	0.7	-0.0067	-1.0	0.0282	92.4
1000	ω_0	-1.0	-0.0256	2.6	0.3998	94.1
	ω_1	3.1	0.0469	1.5	0.4309	95.9
	ω_2	-1.5	-0.0148	1.0	0.1332	95.1
	ω_3	1.3	0.0177	1.4	0.2424	95.4
	β_0	-1.0	-0.0018	0.2	0.0988	95.0
	β_1	-2.9	0.0032	-0.1	0.0791	95.7
	β_2	1.1	-0.0002	0.0	0.0268	95.4
	β_3	-1.1	-0.0015	0.1	0.0517	94.2
	κ	0.5	0.0019	0.4	0.0825	94.7
	σ	0.7	-0.0032	-0.5	0.0183	94.8
3000	ω_0	-1.0	-0.0188	1.9	0.2314	95.0
	ω_1	3.1	0.0198	0.6	0.2393	95.5
	ω_2	-1.5	0.0017	-0.1	0.0774	94.1
	ω_3	1.3	-0.0039	-0.3	0.1346	95.6
	β_0	-1.0	0.0010	-0.1	0.0586	95.0
	β_1	-2.9	0.0006	0.0	0.0498	94.6
	β_2	1.1	-0.0007	-0.1	0.0159	94.4
	β_3	-1.1	0.0008	-0.1	0.0288	95.2
	κ	0.5	0.0030	0.6	0.0475	94.3
	σ	0.7	-0.0010	-0.1	0.0106	95.0

Cuadro 4.2: Resultados de simulación GGCI donde porcentaje de ceros 20 %

Tamaño de muestra	Parámetro	Valor verdadero	Sesgo	Sesgo relativo (%)	RECM	Cobertura al 95 % (%)
200	ω_0	-2.0	-0.0942	4.7	0.7443	94.2
	ω_1	-3.1	-0.0948	3.1	0.7534	94.1
	ω_2	0.9	0.0391	4.3	0.2225	94.0
	ω_3	-1.8	-0.0558	3.1	0.4610	94.0
	β_0	-1.0	-0.0053	0.5	0.2080	95.5
	β_1	-2.9	0.0038	-0.1	0.2203	95.3
	β_2	1.1	-0.0009	-0.1	0.0637	93.9
	β_3	-1.1	0.0008	-0.1	0.1368	93.4
	κ	0.5	0.0120	2.4	0.2105	96.2
	σ	0.7	-0.0194	-2.8	0.0496	91.2
500	ω_0	-2.0	-0.0300	1.5	0.4375	94.4
	ω_1	-3.1	-0.0186	0.6	0.4218	95.3
	ω_2	0.9	0.0110	1.2	0.1303	94.6
	ω_3	-1.8	-0.0173	1.0	0.2664	94.1
	β_0	-1.0	-0.0063	0.6	0.1605	94.2
	β_1	-2.9	0.0002	0.0	0.1447	94.4
	β_2	1.1	0.0007	0.1	0.0478	94.9
	β_3	-1.1	-0.0016	0.1	0.0865	94.7
	κ	0.5	-0.0017	-0.3	0.1294	95.3
	σ	0.7	-0.0067	-1.0	0.0306	93.4
1000	ω_0	-2.0	-0.0104	0.5	0.3152	95.3
	ω_1	-3.1	-0.0242	0.8	0.3072	94.5
	ω_2	0.9	0.0042	0.5	0.0881	94.8
	ω_3	-1.8	0.0028	-0.2	0.1745	94.9
	β_0	-1.0	-0.0023	0.2	0.1102	94.9
	β_1	-2.9	0.0002	0.0	0.1022	94.9
	β_2	1.1	0.0011	0.1	0.0311	95.5
	β_3	-1.1	-0.0023	0.2	0.0587	94.9
	κ	0.5	0.0020	0.4	0.0876	96.0
	σ	0.7	-0.0033	-0.5	0.0205	93.7
3000	ω_0	-2.0	0.0008	0.0	0.1760	95.2
	ω_1	-3.1	-0.0120	0.4	0.1757	94.3
	ω_2	0.9	0.0019	0.2	0.0488	95.7
	ω_3	-1.8	-0.0064	0.4	0.0981	95.5
	β_0	-1.0	-0.0021	0.2	0.0607	95.9
	β_1	-2.9	0.0019	-0.1	0.0599	95.2
	β_2	1.1	-0.0001	0.0	0.0173	94.7
	β_3	-1.1	0.0020	-0.2	0.0350	94.8
	κ	0.5	0.0001	0.0	0.0518	94.4
	σ	0.7	-0.0016	-0.2	0.0116	94.3

Cuadro 4.3: Resultados de simulación GGCI donde porcentaje de ceros 40 %

Tamaño de muestra	Parámetro	Valor verdadero	Sesgo	Sesgo relativo (%)	RECM	Cobertura al 95 % (%)
200	ω_0	-3.0	-0.1001	3.3	0.7017	94.4
	ω_1	-1.2	-0.0175	1.5	0.5854	93.2
	ω_2	1.5	0.0464	3.1	0.2217	95.0
	ω_3	-2.9	-0.0830	2.9	0.4153	94.5
	β_0	-1.0	0.0018	-0.2	0.3266	93.8
	β_1	-2.9	-0.0234	0.8	0.2835	93.6
	β_2	1.1	-0.0043	-0.4	0.0986	94.1
	β_3	-1.1	0.0088	-0.8	0.2039	93.9
	κ	0.5	0.0523	10.5	0.2733	97.0
	σ	0.7	-0.0332	-4.7	0.0669	88.6
500	ω_0	-3.0	-0.0219	0.7	0.4270	94.0
	ω_1	-1.2	-0.0302	2.5	0.3613	94.5
	ω_2	1.5	0.0141	0.9	0.1328	93.8
	ω_3	-2.9	-0.0220	0.8	0.2465	94.2
	β_0	-1.0	-0.0144	1.4	0.1796	94.9
	β_1	-2.9	0.0084	-0.3	0.1725	93.0
	β_2	1.1	0.0000	0.0	0.0543	94.4
	β_3	-1.1	0.0034	-0.3	0.1260	94.6
	κ	0.5	0.0161	3.2	0.1515	95.3
	σ	0.7	-0.0120	-1.7	0.0353	92.2
1000	ω_0	-3.0	-0.0129	0.4	0.3069	94.5
	ω_1	-1.2	-0.0007	0.1	0.2442	95.5
	ω_2	1.5	0.0042	0.3	0.0939	95.1
	ω_3	-2.9	-0.0057	0.2	0.1758	94.3
	β_0	-1.0	-0.0025	0.2	0.1295	94.9
	β_1	-2.9	-0.0020	0.1	0.1241	93.6
	β_2	1.1	0.0001	0.0	0.0389	94.4
	β_3	-1.1	0.0020	-0.2	0.0786	95.7
	κ	0.5	0.0018	0.4	0.1052	95.0
	σ	0.7	-0.0048	-0.7	0.0239	93.1
3000	ω_0	-3.0	-0.0085	0.3	0.1706	94.8
	ω_1	-1.2	0.0086	-0.7	0.1432	94.2
	ω_2	1.5	0.0027	0.2	0.0528	94.8
	ω_3	-2.9	-0.0045	0.2	0.0975	94.6
	β_0	-1.0	0.0011	-0.1	0.0695	95.3
	β_1	-2.9	-0.0038	0.1	0.0690	95.5
	β_2	1.1	0.0001	0.0	0.0215	95.2
	β_3	-1.1	0.0002	0.0	0.0466	95.3
	κ	0.5	0.0012	0.2	0.0584	94.9
	σ	0.7	-0.0019	-0.3	0.0132	94.7

Capítulo 5

Aplicación

El objetivo de este capítulo es estudiar una aplicación del MCIM a datos reales, que permita comparar los resultados del MCIM y los del modelo alternativo MDP. Los datos son sobre gastos en educación de una muestra de adolescentes del estudio Niños del Milenio del 2009 en el Perú.

5.1. Descripción de la base de datos

La base de datos empleada en esta aplicación proviene del estudio longitudinal Niños del Milenio, conocido internacionalmente como *Young Lives* (Morrow, 2017). El estudio recopila datos que describen el bienestar de una misma muestra de aproximadamente 12,000 niños y niñas de Etiopía, India, Perú y Vietnam, en un período de 15 años (entre 2002 y 2016). En términos generales, los aspectos del bienestar estudiados son educación, salud, empleo, uso del tiempo, sentimientos y actitudes, calidad de la vivienda, acceso a programas sociales, entre otros. El estudio es financiado principalmente por el Departamento de Desarrollo Internacional del Gobierno de Reino Unido.

Las instituciones responsables de la implementación del estudio en el Perú son el Grupo de Análisis para el Desarrollo (Grade) y el Instituto de Investigación Nutricional (INN). Hasta el momento han recopilado datos en 5 rondas o momentos (2002, 2006, 2009, 2013 y 2016). La muestra de la primera ronda estuvo conformada por 2,766 niños y niñas de 20 *clusters* o comunidades. La muestra fue dividida en dos grupos: un cohorte menor de 2,052 niños y niñas que en el 2002 tenían menos de 2 años de edad, y un cohorte mayor de 714 niños y niñas que en su mayoría tenían entre 7 y 8 años de edad.

En Grade (2015) se describe que la selección de la muestra se efectuó en dos niveles. Los *clusters* o distritos fueron seleccionados a partir de un marco muestral conformado por los distritos existentes en el 2002 en el país, excluyendo al 5 % de distritos menos pobres y considerando que los distritos con mayor población tuvieran una mayor probabilidad de ser seleccionados. De las 10 muestras generadas de 20 *clusters*, se seleccionó a una que cumpla con los criterios de diversidad (en términos de grupos étnicos, clima, geografía, densidad poblacional, etc.) y de logística y presupuesto del estudio. Una vez definidos los *cluster*, en

cada una se seleccionó al azar a una zona censal y, en ella, a una manzana (en el caso de zonas urbanas) o un centro poblado (en zonas rurales). Las viviendas de cada manzana o centro fueron visitadas para identificar a las familias con niños o niñas cuyas edades eran de interés del estudio.

El estudio de aplicación identificará las variables que determinan los niveles de gastos en educación de adolescentes participantes de Niños del Milenio en el Perú. Específicamente, de la muestra del cohorte mayor de la tercera ronda (2009), que para entonces la conformaban 678 adolescentes que en su mayoría tenían entre 14 y 15 años de edad; aunque, por presencia de datos *missing* en algunas de las covariables, decidimos trabajar con 661 adolescentes. La base de datos es citada como Boyden (2014) y es extraída del repositorio de datos *UK Data Archive*.

Es importante asegurar que los valores cero registrados de la respuesta signifiquen que las personas no están estudiando a pesar de estar en una edad donde es pertinente hacerlo. Por ello, elegimos la ronda 3 del estudio porque en ese momento los participantes son adolescentes con edad de estudiar, 14 y 15 años de edad. En las siguientes rondas estos mismos participantes tienen 18 o más años, en la cual no necesariamente tienen la obligación de estudiar.

Definimos a la variable respuesta como la suma de los gastos directos e indirectos destinados a la educación de adolescentes, efectuados por algún miembro del hogar del adolescente en los últimos 12 meses (en adelante "Gasto en educación"). Los gastos directos son las matrículas, mensualidades, donaciones a la escuela o aportes a la asociación de padres y madres. Los gastos indirectos son debido a clases particulares, uniformes, libros o útiles escolares. Precisamos que los adolescentes con gasto cero son aquellos que afirman no estar asistiendo a la escuela en lo transcurrido del año de la encuesta. Para los adolescentes que afirman asistir a la escuela, consideramos lo reportado por su cuidador o cuidadora principal cuando se le consulta sobre el gasto en un conjunto de productos relacionados a la educación escolar de los miembros del hogar y cuánto de ese gasto total fue hecho solo para el adolescente de estudio. Las opciones de respuesta de la última pregunta fueron "Nada", "Menos de la mitad", "La mitad", "Más de la mitad pero no todo" y "Todo". Con el fin de tener una aproximación del gasto asumimos que la segunda y la cuarta opción son 25 % y 75 %, respectivamente. Expresamos los gastos en ciento de soles.

Las covariables a considerar en el modelo son:

1. "Índice de vivienda": índice sobre la calidad de la vivienda del adolescente. Es un promedio simple de 4 valores: 3 *dummies* y una variable reescalada. Las *dummies* indican si el material principal es el adecuado en la pared ("ladrillo/concreto" o "bloquetas de concreto/ladrillos superpuestos"), el techo ("concreto/cemento", "calamina/fierro" o "teja") y el piso ("cemento/loseta", "laminado/vinílico", "mármol", "piedra pulida" o "parquet"). La variable reescalada es el número de habitaciones por persona reescalado para que tome valores en $[0, 1]$. El intervalo de valores posibles del índice es $[0, 1]$, tal que un valor más alto indica que se tiene mejor calidad de la vivienda.

2. "Índice de consumo": índice sobre el consumo de bienes durables del hogar del adolescente. Es un promedio simple de un conjunto de *dummies* que indican la propiedad o no de un bien de una determinada lista, donde todos los bienes tienen la misma ponderación, independientemente del valor monetario distinto del bien. La lista está compuesta por máquina de coser, televisión, radio, carro o camión, motocicleta, bicicleta, teléfono, refrigeradora, cocina a gas o eléctrica o solar, ventilador, terma, computadora o laptop, plancha y licuadora. El intervalo de valores posibles del índice es $[0, 1]$, tal que un valor más alto indica que se tiene más bienes de la lista.

3. "Menores que estudian": número de miembros del hogar del adolescente menores de 18 años de edad que estudian en educación formal. No incluye al adolescente.

4. "Sexo": sexo del adolescente, donde "Sexo" $= 1$ si es mujer y "Sexo" $= 0$ si es hombre.

5. "Centro de estudios": tipo de centro de estudios del adolescente, donde "Centro de estudios" $= 1$ si estudia en un centro público (del gobierno central o local) en el 2009 y "Centro de estudios" $= 0$ si estudia en uno de tipo privado. En el caso que el adolescente no esté estudiando, registramos el tipo del centro de estudios que por última vez asistió.

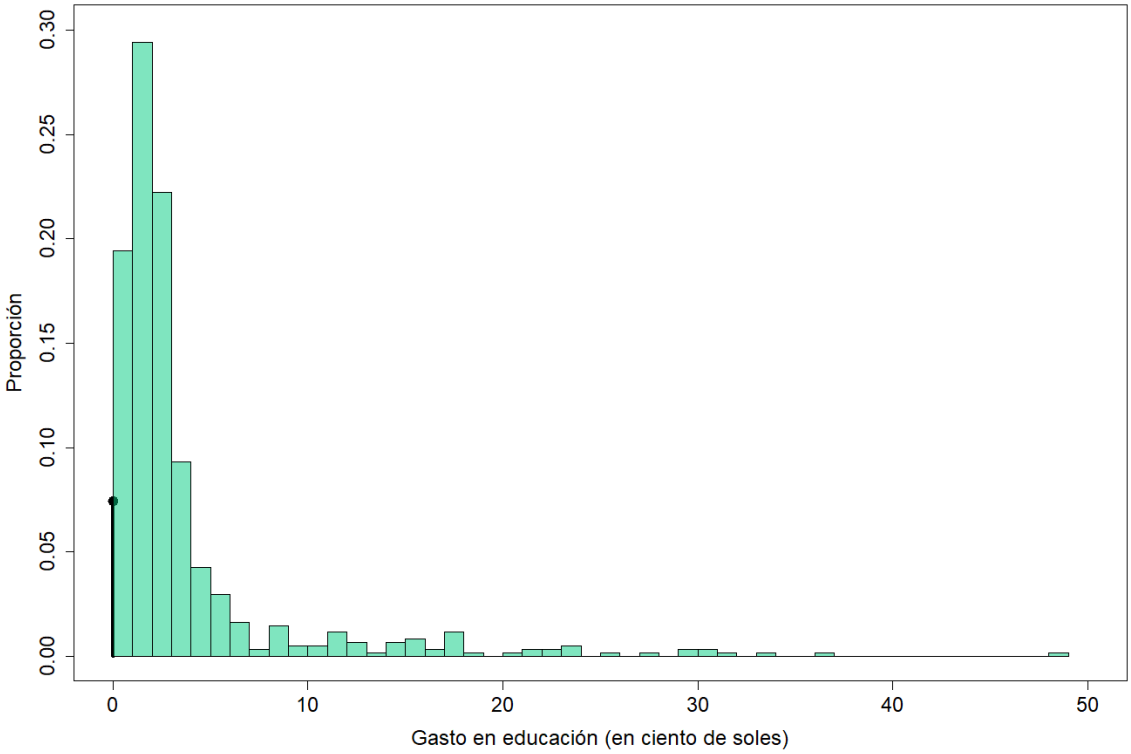
6. "Años de educación": número de años de educación formal (en los niveles de educación primaria o secundaria) que el adolescente ha culminado satisfactoriamente hasta el año 2009.

5.2. Estadísticas descriptivas

La distribución empírica de los gastos en educación de la muestra de adolescentes, mostrada en la Figura 5.1, tiene asimetría positiva, presentando una alta concentración en valores pequeños: aproximadamente el 85 % de los adolescentes tienen gastos en educación menores de 500 soles anuales, siendo la media de 342 soles mayor a la mediana de 185 soles. Además, la distribución tiene una cola ancha, donde 6 adolescentes registran gastos mayores de 3,000 soles y el valor máximo registrado es 4,900 soles. A diferencia de las columnas del histograma, tenemos un bastón en el nivel cero del eje horizontal cuya altura señala que aproximadamente el 7 % de los adolescentes afirman no estar estudiando y, por consiguiente, sus hogares no efectúan gasto en educación para el adolescente.

Existen diferencias entre los adolescentes que estudian (con gasto positivo) y los que no (con gasto cero), como ilustra el Cuadro 5.1. La más notoria es la diferencia de la vivienda y el consumo, llevando a proponer que es un factor en la decisión sobre gastar o no en educación. Los respectivos promedios del "Índice de vivienda" y del "Índice de consumo" en los que no estudian son menores comparado con los que estudian, siendo además sus desviaciones estándar menores en el primer grupo, indicando que es un grupo muy parecido en términos de estas covariables y, por ende, el menor promedio puede estar representándolos mejor. Entre los que no estudian, el 92 % lo hizo por última vez en un centro público, el 35 % es mujer, y pudo culminar alrededor de 6 años de educación formal (el equivalente a tener una educación primaria completa); siendo estos indicadores distantes entre los adolescentes que sí estudian.

Figura 5.1: Histograma de "Gasto en educación"

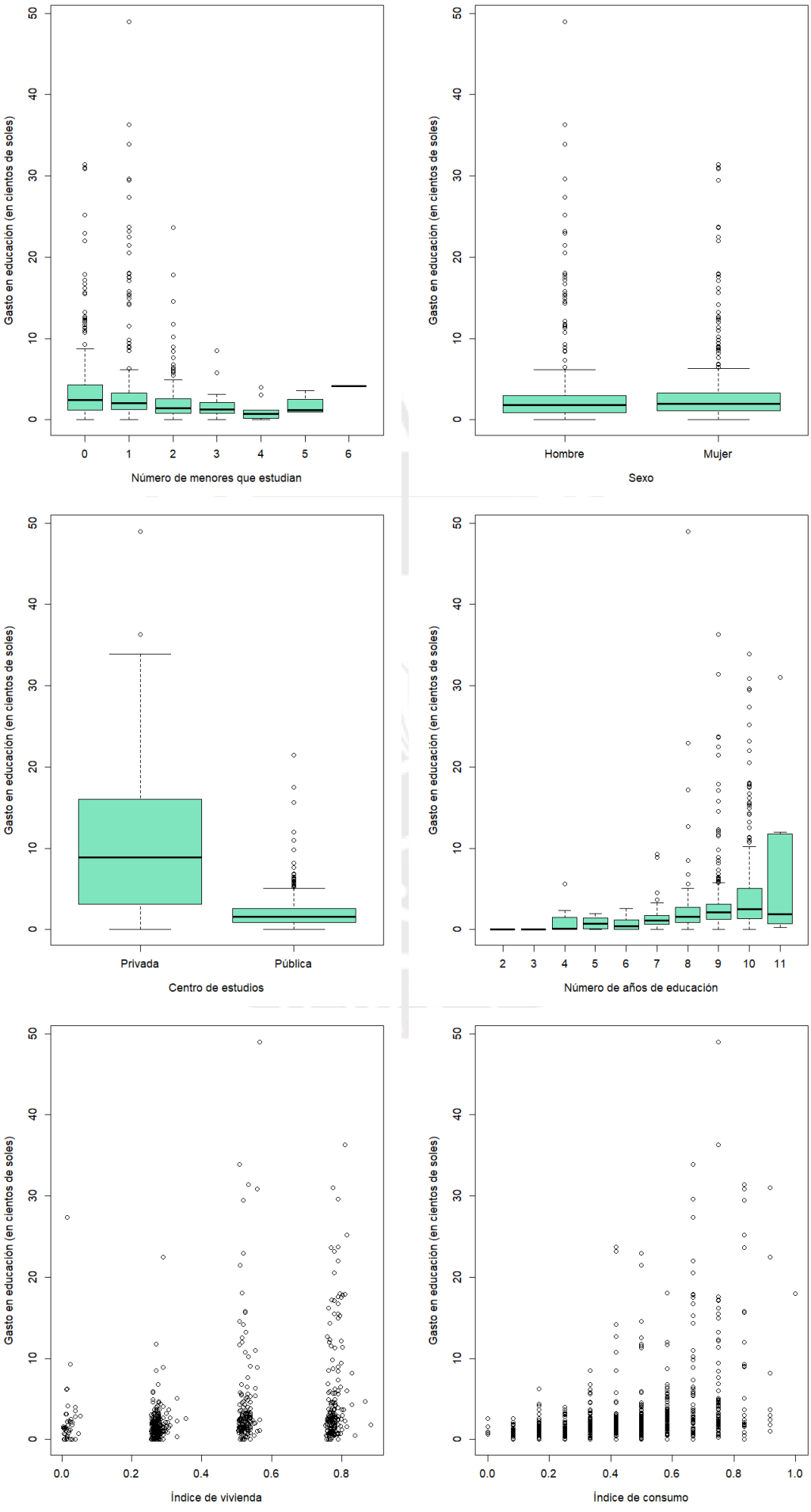


Cuadro 5.1: Características de los adolescentes según decisión de gastar

Covariables	No efectúan gastos en educación n=49	Sí efectúan gastos en educación n=612	Total n=661
Menores que estudian (media; desv)	1.32; 1.21	1.19; 1.06	1.20; 1.07
Índice de vivienda (media; desv)	0.3139; 0.2094	0.4721; 0.2357	0.4604; 0.2367
Índice de consumo (media; desv)	0.3112; 0.2039	0.4764; 0.2136	0.4641; 0.2174
Sexo: mujer (n;%)	17 (35 %)	291 (48 %)	308 (47 %)
Centro de estudios: pública (n;%)	45 (92 %)	513 (84 %)	558 (84 %)
Años de educación (media; desv)	6.32; 1.65	8.82; 1.22	8.64; 1.39

Los gráficos de cajas y de dispersión de la Figura 5.2 proponen la inclusión de algunas características como covariables en el modelo de regresión. A raíz de ellos un proposición es que la respuesta se asocia fuertemente con "Centro de estudios", "Años de educación" e "Índice de vivienda". El gasto entre los que estudian en un centro público está muy concentrado en gastos menores de 400 soles aproximadamente, pero el gasto en centros privados está más disperso y de mayor mediana. La asociación entre "Años de educación" y la respuesta puede entenderse con el aumento de la mediana del gasto conforme los adolescentes culminan niveles de educación formal. Existiría aparentemente una influencia menos notoria de "Menores que estudian" y "Sexo" sobre el gasto en educación.

Figura 5.2: Gráficos de cajas y dispersión de "Gasto en educación" según covariables



5.3. Especificación del modelo

En el MCIM-GG suponemos que la respuesta "Gasto en educación" para el adolescente i tiene distribución $Y_i \sim GGCI(\delta_i, \gamma_i, \kappa, \sigma)$ en términos de (2.37) y que los parámetros $\delta_i = P(Y_i = 0)$ y $\gamma_i = E(Y_i)$ están relacionados con determinadas covariables:

$$Y_i \sim GGCI(\delta_i, \gamma_i, \kappa, \sigma) \tag{5.1}$$

$$\text{logit}(\delta_i) = \omega_0 + \omega_1\, x_{1i} + \omega_2\, x_{2i} + \omega_3\, x_{3i} \tag{5.2}$$

$$\gamma_i = \exp\left(\beta_0 + \beta_1\, x_{1i} + \beta_2\, x_{2i} + \beta_4\, x_{4i} + \beta_5\, x_{5i} + \beta_6\, x_{6i}\right) \tag{5.3}$$

donde x_{1i} es el "Índice de vivienda" y x_{2i} es el "Índice de consumo", ambas son covariables incluidas en las dos ecuaciones de regresión. La covariable x_{3i} es "Menores que estudian" y está únicamente en la primera ecuación. Las demás covariables, x_{4i} es "Sexo", x_{5i} es "Centro de estudios" y x_{6i} es "Años en educación", están únicamente en la segunda ecuación.

En el MDP-GG suponemos que la respuesta tiene distribución $Y_i \sim GGCI(\delta_i, \mu_i, \kappa, \sigma)$ en términos de (2.32) y que los parámetros $\delta_i = P(Y_i = 0)$ y $\mu_i = E(Y_i\,|Y_i > 0)$ están relacionados con covariables:

$$Y_i \sim GGCI(\delta_i, \mu_i, \kappa, \sigma) \tag{5.4}$$

$$\text{logit}(\delta_i) = \omega_0 + \omega_1\, x_{1i} + \omega_2\, x_{2i} + \omega_3\, x_{3i} \tag{5.5}$$

$$\mu_i = \exp\left(\beta_0 + \beta_1\, x_{1i} + \beta_2\, x_{2i} + \beta_4\, x_{4i} + \beta_5\, x_{5i} + \beta_6\, x_{6i}\right) \tag{5.6}$$

Los parámetros κ y σ son considerados constantes en todas las observaciones.

La selección de las covariables del MCIM-GG y MDP-GG estuvo en primer lugar en función de obtener criterios de información adecuados y estimaciones estadísticamente significativas de los coeficientes de regresión, procurando que ambos modelos tengan las mismas covariables o especificaciones de sus ecuaciones de regresiones.

La selección también estuvo en función de explorar relaciones de interés como, por ejemplo, entre el valor esperado del gasto (parámetro γ_i) y la covariable "Sexo", y confirmar la existencia de algún tipo de discriminación en contra de mujeres. Además, estudiar si la decisión del hogar o del adolescente de que éste último estudie o no (el parámetro δ_i) está determinado únicamente por indicadores digamos materiales del hogar ("Índice de vivienda" y "Índice de consumo") y por el número de "Menores que estudian". Por último, "Años de educación", para explorar la carga del gasto para el hogar en la medida que el adolescente avanza en su educación escolar, así como, para quienes no estudian, conocer si el rezago educativo de un adolescente conlleva a que el hogar le asigne un menor apoyo, un menor gasto.

Para la estimación del MCIM-GG utilizamos un código en Matlab basado en la función "fmincon", descrito en el Apéndice B.6. Mientras que para la estimación del MDP-GG utilizamos un código en SAS, el cual es una adaptación al presentado por Smith y Preisser (2018), descrito en el Apéndice C.

5.4. Resultados

En los Cuadros 5.2 y 5.3 mostramos las estimaciones de los parámetros de regresión de los modelos MCIM-GG y MDP-GG, respectivamente; así como sus errores estándar basados en la matriz de información de Fisher observada evaluada en las correspondientes estimaciones, la significancia individual de los parámetros según la estadística de Wald y la exponencial de las estimaciones.

En el MCIM-GG, las exponenciales de los parámetros de regresión β_j pueden ser interpretadas como los efectos multiplicativos del incremento de una unidad de la correspondiente covariable, x_{ji}, sobre el valor esperado o la media de la respuesta, $E(Y_i)$ (el efecto multiplicativo sobre la estimación del parámetro γ_i). El efecto multiplicativo de, por ejemplo, x_{2i} está dado por:

$$\frac{E(Y_i|x_{2i} = a + 1\,\%)}{E(Y_i|x_{2i} = a)} = \exp(\beta_2) \tag{5.7}$$

donde a es cualquier valor dado. Recordar que x_{2i} es un índice que toma valores entre 0 y 1, por tanto, consideramos que el incremento de una unidad es un incremento en 1 %.

Por ejemplo, el aumento marginal de 1 en "Años de educación" de un adolescente (es decir, tener un año más de educación formal culminado satisfactoriamente) conllevaría a un aumento en 6.8 % del valor esperado de "Gasto en educación".

En cambio, en el MDP-GG, las exponenciales de los parámetros de regresión β_j miden un efecto multiplicativo distinto, miden el efecto sobre el valor esperado o la media de la respuesta condicionada a que el adolescente está estudiando (el efecto multiplicativo sobre la estimación del parámetro μ_i). Por ejemplo, ahora un aumento marginal de 1 en "Años de educación" de un adolescente conllevaría a un aumento en 7.2 % del valor esperado de "Gasto en educación" condicionado a que el adolescente estudia (condicionado a que el gasto es positivo).

El MDP-GG tiene dificultades para estimar un efecto multiplicativo de una covariable sobre la media total, $\gamma_i = (1 - \delta_i)\,\mu_i$, debido a que esta estimación se construiría a partir de dos ecuaciones de regresión no lineales. Por ejemplo, el efecto multiplicativo de la covariable x_{2i}, que por cierto es una covariable incluida en las dos ecuaciones del MDP-GG, está dado por:

$$\frac{E(Y_i|x_{2i} = a + 1\,\%)}{E(Y_i|x_{2i} = a)} = \frac{1 + \exp(\omega_0 + \omega_1\, x_{1i} + \omega_2\, a + \omega_3 x_{3i})}{1 + \exp(\omega_0 + \omega_1\, x_{1i} + \omega_2\, (a + 1\,\%) + \omega_3 x_{3i})}\, \exp(\beta_2) \tag{5.8}$$

Observamos que el MDP-GG puede tener una aproximación de este efecto pero restringida a valores específicos de las covariables incluidas en la ecuación de δ_i. En cambio, en el modelo MCIM-GG, como está descrito, las exponenciales de las estimaciones pueden ser interpretadas directamente, sin necesidad de especificar valores de otras covariables.

Otro apunte es que las covariables que son estadísticamente significativas de forma individual al 5 % en un modelo también lo son en el otro. Aunque observamos que, en el modelo

MCIM-GG, el coeficiente de "Menores que estudian" de la ecuación de δ_i tiene un p-valor de 0.052; sin embargo, en el modelo MDP-GG, este mismo coeficiente es relativamente menor en valor absoluto y tiene un p-valor de 0.473. Los coeficientes de "Índice de consumo" y "Años de educación" de la ecuación γ_i del MCIM-GG (de la ecuación μ en el caso de MDP-GG) son positivos y estadísticamente significativos. Solamente "Centro de estudios" tiene un coeficiente negativo estadísticamente significativo, lo que es esperable porque el gasto se aminora si el adolescente estudia en un centro público, donde la educación es gratuita. No existe evidencia que el "Sexo" influya sobre la media total del gasto en educación.

Finalmente, de acuerdo a los criterios AIC, AICc y BIC, que el Cuadro 5.4 muestra para ambos modelos, verificamos que el MCIM-GG de esta aplicación tiene un óptimo ajuste respecto al MDP-GG.

Cuadro 5.2: Estimación de coeficientes de regresión del MCIM-GG

Parámetro	Covariable	Coeficiente de regresión estimado	Error estándar	P-valor	Exponencial del coeficiente
δ_i	Intercepto	-0.189	0.384	0.624	0.828
	Índice de vivienda (x_{1i})	-2.140	0.816	0.009	0.118
	Índice de consumo (x_{2i})	-2.979	0.817	<0.001	0.051
	Menores que estudian (x_{3i})	-0.257	0.132	0.052	0.774
γ_i	Intercepto	0.630	0.251	0.012	1.878
	Índice de vivienda (x_{1i})	0.244	0.154	0.112	1.277
	Índice de consumo (x_{2i})	1.476	0.172	<0.001	4.373
	Sexo (x_{4i})	0.051	0.057	0.370	1.053
	Centro de estudios (x_{5i})	-1.342	0.084	<0.001	0.261
	Años de educación (x_{6i})	0.065	0.027	0.015	1.068
κ		0.554	0.077	<0.001	
σ		0.702	0.021	<0.001	

Cuadro 5.3: Estimación de coeficientes de regresión del MDP-GG

Parámetro	Covariable	Coeficiente de regresión estimado	Error estándar	P-valor	Exponencial del coeficiente
δ_i	Intercepto	-0.490	0.415	0.239	0.613
	Índice de vivienda (x_{1i})	-2.031	0.836	0.015	0.131
	Índice de consumo (x_{2i})	-2.888	0.868	<0.001	0.056
	Menores que estudian (x_{3i})	-0.096	0.134	0.473	0.908
μ_i	Intercepto	0.834	0.250	<0.001	2.302
	Índice de vivienda (x_{1i})	0.112	0.144	0.438	1.119
	Índice de consumo (x_{2i})	1.273	0.162	<0.001	3.572
	Sexo (x_{4i})	0.057	0.058	0.323	1.059
	Centro de estudios (x_{5i})	-1.346	0.084	<0.001	0.260
	Años de educación (x_{6i})	0.069	0.027	0.010	1.072
κ		0.548	0.078	<0.001	
σ		0.706	0.022	<0.001	

Cuadro 5.4: Criterios de información de los modelos MCIM-GG y MDP-GG

Criterios	MCIM-GG	MDP-GG
Log verosimilitud	-1,301.4	-1,303.4
AIC	2,626.8	2,630.8
AICc	2,627.3	2,631.3
BIC	2,680.8	2,684.7
RECM	3.8413	3.9543

Capítulo 6

Conclusiones

6.1. Conclusiones

Hemos desarrollado un modelo de regresión cero-inflacionada a la media (MCIM) de una respuesta semicontinua, como alternativa al modelo de regresión de dos partes (MDP). Con el MCIM es factible estimar e interpretar los efectos de un conjunto de covariables sobre la media total de la respuesta, en vez de la media condicionada a valores positivos, y sobre la probabilidad que esta respuesta tome el valor cero. Además, estudiamos el MCIM suponiendo que los valores positivos de la respuesta presentan una distribución gamma generalizada (MCIM-GG). Mediante un estudio de simulación, hemos observado que las estimaciones de máxima verosimilitud del MCIM-GG tienen un adecuado desempeño en términos de sesgo, sesgo relativo, raíz de error cuadrático medio y cobertura al 95 % del intervalo de confianza. Además, desarrollamos un estudio de aplicación del MCIM-GG y del MDP-GG a datos de gastos en educación de una muestra de adolescentes del estudio Niños del Milenio del 2009 en el Perú, hallando escenarios donde el MCIM-GG tiene un mejor ajuste a los datos respecto al MDP-GG, así como interpretaciones más directas de efectos de covariables sobre la media del gasto.

6.2. Sugerencias para investigaciones futuras

Sugerimos como investigaciones futuras a las siguientes:

1. Estudiar el MCIM-GG con una especificación que incluya una ecuación de regresión para un parámetro de dispersión de la distribución continua positiva del modelo.
2. Estudiar el MCIM-GG con una especificación que incluya funciones de enlace distintas a la logarítmica o la logística.
3. Estudiar la estimación bayesiana del MCIM-GG.
4. Estudiar el MCIM-GG en datos longitudinales.
5. Aplicar el MCIM-GG a datos de gastos provenientes de encuestas de hogares como, por ejemplo, la ENAHO.

Apéndice A

Resultados teóricos

A.1. Reexpresión de la función de densidad gamma generalizada

De acuerdo a Stacy y Mihram (1965), la función de densidad de la distribución GG de parámetros $\alpha > 0$, $\beta > 0$ y $\rho \neq 0$ está dada por:

$$gg(y \mid \alpha, \beta, \rho) = \frac{1}{\Gamma(\alpha)} \left(\frac{1}{\beta}\right)^{\alpha\rho} |\rho| y^{\alpha\rho-1} \exp\left(-\left(\frac{y}{\beta}\right)^{\rho}\right) \quad , y \geq 0,$$

Ahora, la parametrización propuesta en Manning et al. (2005) establece que α, β y ρ sean expresados en función de nuevos parámetros λ, κ y σ de la siguiente manera:

$$\begin{aligned} \alpha &= \frac{1}{|\kappa|^2} \\ \beta &= \frac{\exp(\lambda)}{|\kappa|^{-2 \text{ Signo}(\kappa)\, \sigma/|\kappa|}} \\ \rho &= \text{Signo}(\kappa)\frac{|\kappa|}{\sigma} \end{aligned}$$

El procedimiento para hallar la nueva expresión de la función de densidad (2.22) es detallado a continuación:

$$\begin{aligned} gg(y \mid \lambda, \kappa, \sigma) &= \frac{1}{\Gamma(1/|\kappa|^2)} \left(\frac{|\kappa|^{-2 \text{ Signo}(\kappa)\, \frac{\sigma}{|\kappa|}}}{\exp(\lambda)}\right)^{\frac{1}{|\kappa|^2}\text{Signo}(\kappa)\frac{|\kappa|}{\sigma}} \left|\text{Signo}(\kappa)\frac{|\kappa|}{\sigma}\right| y^{\frac{1}{|\kappa|^2}\text{Signo}(\kappa)\frac{|\kappa|}{\sigma}-1} \\ &\quad \exp\left(-\left(y\, \frac{|\kappa|^{-2 \text{ Signo}(\kappa)\frac{\sigma}{|\kappa|}}}{\exp(\lambda)}\right)^{\text{Signo}(\kappa)\frac{|\kappa|}{\sigma}}\right) \\ &= \frac{1}{\Gamma(1/|\kappa|^2)}\, \frac{|\kappa|^{\frac{-2}{|\kappa|^2}}}{\exp(\frac{1}{|\kappa|^2}\text{Signo}(\kappa)\frac{|\kappa|}{\sigma}\lambda)}\, \frac{|\kappa|}{\sigma}\, \frac{1}{y} y^{\frac{1}{|\kappa|^2}\text{Signo}(\kappa)\frac{|\kappa|}{\sigma}} \\ &\quad \exp\left(-y^{\text{Signo}(\kappa)\frac{|\kappa|}{\sigma}} \frac{|\kappa|^{-2}}{\exp\left(\text{Signo}(\kappa)\frac{|\kappa|}{\sigma}\lambda\right)}\right) \end{aligned}$$

$$
\begin{aligned}
gg(y\,|\,\lambda,\kappa,\sigma) &= \frac{1}{\Gamma(1/|\kappa|^2)}|\kappa|^{\frac{-2}{|\kappa|^2}}\frac{|\kappa|}{\sigma}\frac{1}{y}\exp\left(-\frac{1}{|\kappa|^2}\text{Signo}(\kappa)\frac{|\kappa|}{\sigma}\lambda\right)\ \exp\left(\frac{1}{|\kappa|^2}\text{Signo}(\kappa)\frac{|\kappa|}{\sigma}\ln(y)\right)\\
&\quad \exp\left(-\frac{\exp\left(\text{Signo}(\kappa)\frac{|\kappa|}{\sigma}\ln(y)\right)}{\exp\left(\text{Signo}(\kappa)\frac{|\kappa|}{\sigma}\lambda\right)}|\kappa|^{-2}\right)\\
&= \frac{|\kappa|^{-2/|\kappa|^2}}{\sigma y|\kappa|^{-1}\Gamma(1/|\kappa|^2)}\ \exp\left(\frac{1}{|\kappa|}\text{Signo}(\kappa)\frac{\ln(y)-\lambda}{\sigma}\right)\\
&\quad \exp\left(-\frac{1}{|\kappa|^2}\exp\left(|\kappa|\text{Signo}(\kappa)\frac{\ln(y)-\lambda}{\sigma}\right)\right)\\
&= \frac{|\kappa|^{-2/|\kappa|^2}}{\sigma y|\kappa|^{-1}\Gamma(1/|\kappa|^2)}\exp\left(\frac{\nu}{|\kappa|}-\frac{\exp(|\kappa|\nu)}{|\kappa|^2}\right)
\end{aligned}
$$

donde $\nu = \text{Signo}(\kappa)\dfrac{\ln(y)-\lambda}{\sigma}$

Entonces, una variable aleatoria continua Y presenta distribución gamma generalizad de parámetros $-\infty < \lambda < \infty$, $\kappa \neq 0$ y $\sigma > 0$ si su función de densidad está dada por:

$$
gg(y\,|\,\lambda,\kappa,\sigma) = \frac{|\kappa|^{-2/|\kappa|^2}}{\sigma y|\kappa|^{-1}\Gamma(1/|\kappa|^2)}\exp\left(\frac{\nu}{|\kappa|}-\frac{\exp(|\kappa|\nu)}{|\kappa|^2}\right), \qquad y \geq 0, \tag{A.1}
$$

donde $\nu = \text{Signo}(\kappa)\dfrac{\ln(y)-\lambda}{\sigma}$.

Considerando la gamma generalizada con un espacio paramétrico de κ restringido a $\kappa > 0$, volvemos a hacer el procedimiento anterior para obtener una nueva expresión que ayudará a ordenar los códigos en Matlab y la obtención de primeras y segundas derivadas de la función de log-verosimilitud. El resultado final será la expresión de 2.23:

$$
\begin{aligned}
gg(y\,|\lambda,\kappa,\sigma) &= \frac{1}{\Gamma(1/\kappa^2)}\left(\frac{\kappa^{-\frac{2\sigma}{\kappa}}}{\exp(\lambda)}\right)^{\frac{1}{\kappa^2}\frac{\kappa}{\sigma}}\frac{\kappa}{\sigma}y^{\frac{1}{\kappa^2}\frac{\kappa}{\sigma}-1}\exp\left(-y^{\frac{\kappa}{\sigma}}\left(\frac{\kappa^{-2\sigma/\kappa}}{\exp(\lambda)}\right)^{\frac{\kappa}{\sigma}}\right)\\
&= \frac{1}{\Gamma(1/\kappa^2)}\frac{\kappa^{-\frac{2}{\kappa^2}}}{\exp(\lambda/\kappa\sigma)}\frac{\kappa}{\sigma}y^{\frac{1}{\kappa\sigma}-1}\exp\left(-y^{\frac{\kappa}{\sigma}}\frac{\kappa^{-2}}{\exp(\lambda\kappa/\sigma)}\right)\\
&= \frac{\kappa/\sigma}{\Gamma(1/\kappa^2)}y^{\frac{1}{\kappa\sigma}-1}\exp\left(\ln\left(\kappa^{-\frac{2}{\kappa^2}}\right)\right)\exp\left(-\frac{\lambda}{\kappa\sigma}\right)\exp\left(-\frac{1}{\kappa^2}y^{\frac{\kappa}{\sigma}}\exp\left(-\frac{\kappa\lambda}{\sigma}\right)\right)\\
&= \frac{\kappa/\sigma}{\Gamma(1/\kappa^2)}y^{\frac{1}{\kappa\sigma}-1}\exp\left(\ln\left(\kappa^{-\frac{2}{\kappa^2}}\right)-\frac{\lambda}{\kappa\sigma}-\frac{1}{\kappa^2}y^{\frac{\kappa}{\sigma}}\exp\left(-\frac{\kappa\lambda}{\sigma}\right)\right)\\
&= \frac{\kappa/\sigma}{\Gamma(1/\kappa^2)}y^{\frac{1}{\kappa\sigma}-1}\exp\left(\frac{1}{\kappa^2}\ln\left(\frac{1}{\kappa^2}\right)+\frac{1}{\kappa^2}\ln\exp\left(-\frac{\kappa\lambda}{\sigma}\right)-\frac{1}{\kappa^2}y^{\frac{\kappa}{\sigma}}\exp\left(-\frac{\kappa\lambda}{\sigma}\right)\right)\\
&= \frac{\kappa/\sigma}{\Gamma(1/\kappa^2)}y^{\frac{1}{\kappa\sigma}-1}\exp\left(\frac{1}{\kappa^2}\ln\left(\frac{1}{\kappa^2}\exp\left(-\frac{\kappa\lambda}{\sigma}\right)\right)-\frac{1}{\kappa^2}y^{\frac{\kappa}{\sigma}}\exp\left(-\frac{\kappa\lambda}{\sigma}\right)\right)\\
&= \frac{\kappa/\sigma}{\Gamma(1/\kappa^2)}y^{\frac{1}{\kappa\sigma}-1}\exp\left(-\frac{\eta}{\kappa^2}y^{\frac{\kappa}{\sigma}}+\frac{1}{\kappa^2}\ln\left(\frac{\eta}{\kappa^2}\right)\right)
\end{aligned}
$$

donde $\eta = \exp\left(-\dfrac{\kappa\lambda}{\sigma}\right)$.

A.2. Primeras y segundas derivadas de la función de verosimilitud del MCIM-GG

En esta parte desarrollamos las derivadas de primer y segundo orden de la función log-verosimilitud del MCIM-GG 3.13.

Previamente, se detallará algunas derivadas básicas que utilizaremos para obtener las primera y segundas derivadas. Primero, partiendo de las funciones de enlace g_1 y g_2, detalladas en las Ecuaciones (3.7) y (3.8), respectivamente, se obtiene las siguientes derivadas:

$$\delta_i = g_1^{-1}(\widetilde{\boldsymbol{x}_i}^{\boldsymbol{t}}\boldsymbol{\omega}) = \frac{1}{1+\exp(-\widetilde{\boldsymbol{x}_i}^{\boldsymbol{t}}\boldsymbol{\omega})}$$
$$\frac{\partial \delta_i}{\partial \omega_j} = \frac{1}{g_1'(\delta_i)}\widetilde{x}_{ij} = \delta_i(1-\delta_i)\widetilde{x}_{ij}$$

$$\gamma_i = g_2^{-1}(\boldsymbol{x}_i^{\boldsymbol{t}}\boldsymbol{\beta}) = \exp(\boldsymbol{x}_i^{\boldsymbol{t}}\boldsymbol{\beta})$$
$$\frac{\partial \gamma_i}{\partial \beta_j} = \frac{1}{g_2'(\gamma_i)}x_{ij} = \gamma_i x_{ij}$$

Segundo, partiendo de la expresión del λ_i descrita en la Ecuación 3.9, se obtiene las siguientes derivadas:

$$\lambda_i = \ln\gamma_i - \ln(1-\delta_i) - \frac{2\sigma}{k}\ln k + \ln\Gamma(\frac{1}{k^2}) - \ln\Gamma(\frac{1}{k^2}+\frac{\sigma}{k})$$

$$\frac{\partial \lambda_i}{\partial \gamma_i} = \frac{1}{\gamma_i}$$
$$\frac{\partial \lambda_i}{\partial \delta_i} = \frac{1}{1-\delta_i}$$
$$\frac{\partial \lambda_i}{\partial k} = -\frac{2\sigma}{k^2}(1-\ln k) - \frac{2}{k^3}\psi(\frac{1}{k^2}) + (\frac{2}{k^3}+\frac{\sigma}{k^2})\psi(\frac{1}{k^2}+\frac{\sigma}{k})$$
$$\frac{\partial \lambda_i}{\partial \sigma} = -\frac{2}{k}\ln k - \frac{1}{k}\psi(\frac{1}{k^2}+\frac{\sigma}{k})$$
$$\frac{\partial \lambda_i}{\partial \beta_j} = x_{ij}$$

$$\frac{\partial^2 \lambda_i}{\partial \kappa^2} = \frac{2\sigma}{\kappa^3} + \frac{4\sigma}{\kappa^3}(1-\ln\kappa) + \frac{4}{\kappa^6}\frac{\partial\psi(1/\kappa^2)}{\partial(1/\kappa^2)} + \frac{6}{\kappa^4}\psi(1/\kappa^2) - (\frac{2}{\kappa^3}+\frac{\sigma}{\kappa^2})^2\frac{\partial\psi(1/\kappa^2+\sigma/\kappa)}{\partial(1/\kappa^2+\sigma/\kappa)}$$
$$-(\frac{6}{\kappa^4}+\frac{2\sigma}{\kappa^3})\psi(1/\kappa^2+\sigma/\kappa)$$
$$\frac{\partial^2 \lambda_i}{\partial \sigma^2} = -\frac{1}{\kappa^2}\frac{\partial\psi(1/\kappa^2+\sigma/\kappa)}{\partial(1/\kappa^2+\sigma/\kappa)}$$
$$\frac{\partial^2 \lambda_i}{\partial \kappa \partial \sigma} = -\frac{2}{\kappa^2}(1-\ln\kappa) + (\frac{2}{\kappa^4}+\frac{\sigma}{\kappa^3})\frac{\partial\psi(1/\kappa^2+\sigma/\kappa)}{\partial(1/\kappa^2+\sigma/\kappa))} + \frac{1}{\kappa^2}\psi(1/\kappa^2+\sigma/\kappa)$$

donde ψ es la función digamma definida como $\psi(a) = \dfrac{\partial \ln\Gamma(a)}{\partial a}$ para todo $a > 0$.

Además, partiendo de la expresión de η_i descrita en la Ecuación 3.9, se obtiene las si-

guientes derivadas:

$$\eta_i = \exp(-\frac{k\lambda_i}{\sigma})$$

$$\frac{\partial \eta_i}{\partial \lambda_i} = -\eta_i \frac{k}{\sigma}$$
$$\frac{\partial \eta_i}{\partial k} = -\eta_i(\frac{k}{\sigma}\frac{\partial \lambda_i}{\partial k} + \frac{\lambda_i}{\sigma})$$
$$\frac{\partial \eta_i}{\partial \sigma} = -\eta_i(\frac{k}{\sigma}\frac{\partial \lambda_i}{\partial \sigma} - \frac{\lambda_i k}{\sigma^2})$$

$$\frac{\partial^2 \eta_i}{\partial \kappa^2} = -\frac{\kappa}{\sigma}\eta_i\frac{\partial^2 \lambda_i}{\partial \kappa^2} - \frac{\kappa}{\sigma}\frac{\partial \lambda_i}{\partial \kappa}\frac{\partial \eta_i}{\partial \kappa} - \frac{2}{\sigma}\eta_i\frac{\partial \lambda_i}{\partial \kappa} - \frac{1}{\sigma}\lambda_i\frac{\partial \eta_i}{\partial \kappa}$$
$$\frac{\partial^2 \eta_i}{\partial \sigma^2} = -\frac{\kappa}{\sigma}\eta_i\frac{\partial^2 \lambda_i}{\partial \sigma^2} - \frac{\kappa}{\sigma}\frac{\partial \lambda_i}{\partial \sigma}\frac{\partial \eta_i}{\partial \sigma} + \frac{2\kappa}{\sigma^2}\eta_i\frac{\partial \lambda_i}{\partial \sigma} + \frac{\kappa}{\sigma^2}\eta_i\frac{\partial \eta_i}{\partial \sigma} - \frac{2\kappa}{\sigma^3}\lambda_i\eta_i$$
$$\frac{\partial^2 \eta_i}{\partial \kappa \partial \sigma} = \frac{\kappa}{\sigma^2}\lambda_i\frac{\partial \eta_i}{\partial \kappa} + \frac{\kappa}{\sigma^2}\eta_i\frac{\partial \lambda_i}{\partial \kappa} + \frac{1}{\sigma^2}\lambda_i\eta_i - \frac{\kappa}{\sigma}\eta_i\frac{\partial^2 \lambda_i}{\partial \kappa \partial \sigma} - \frac{\kappa}{\sigma}\frac{\partial \lambda_i}{\partial \sigma}\frac{\partial \eta_i}{\partial \kappa} - \frac{1}{\sigma}\eta_i\frac{\partial \lambda_i}{\partial \sigma}$$

Además, se crea una función auxiliar S_i dado que es una expresión que está presente en muchas derivadas de la log-verosimilitud y, por tanto, ayudará a presentar mejor las expresiones de las mismas:

$$S_i = y_i^{\kappa/\sigma}\eta_i$$

$$\frac{\partial S_i}{\partial \sigma} = \frac{\kappa}{\sigma^2}(\lambda_i - \sigma\frac{\partial \lambda_i}{\partial \sigma} - \ln y_i)S_i$$
$$\frac{\partial S_i}{\partial \kappa} = -\frac{1}{\sigma}(\lambda_i + \kappa\frac{\partial \lambda_i}{\partial \kappa} - \ln y_i)S_i$$
$$\frac{\partial S_i}{\partial \beta_j} = -\frac{\kappa}{\sigma}S_i x_{ij}$$

Entonces, las derivadas de primer orden de la log-verosimilitud del MCIM-GG están dadas por:

$$\frac{\partial \mathbb{K}(\theta)}{\partial \omega_j} = \sum_{i=1}^{n_0}(1-\delta_i)\widetilde{x}_{ij} + \sum_{i=n_0+1}^{n} -\delta_i\widetilde{x}_{ij} + \sum_{i=n_0+1}^{n} -\frac{1}{\kappa\sigma}(1-S_i)\delta_i\widetilde{x}_{ij}$$

$$\frac{\partial \mathbb{K}(\theta)}{\partial \beta_j} = \sum_{i=n_0+1}^{n} -\frac{1}{\kappa\sigma}(1-S_i)x_{ij}$$

$$\frac{\partial \mathbb{K}(\theta)}{\partial \kappa} = \sum_{i=n_0+1}^{n} \frac{1}{\kappa} - \frac{1}{\kappa^2\sigma}\ln y_i + \frac{2}{\kappa^3}\psi(1/\kappa^2) + \frac{2}{\kappa^3}S_i - \frac{1}{\kappa^2}\frac{\partial S_i}{\partial \kappa} - \frac{1}{\kappa^2\sigma}(\kappa\frac{\partial \lambda_i}{\partial \kappa} + \lambda_i) + \frac{2\lambda_i}{\kappa^2\sigma} - \frac{2}{\kappa^3} + \frac{2}{\kappa^3}\ln\kappa^2$$

$$\frac{\partial \mathbb{K}(\theta)}{\partial \sigma} = \sum_{i=n_0+1}^{n} -\frac{1}{\sigma} - \frac{1}{\kappa\sigma^2}\ln y_i(1-S_i) - \frac{1}{\kappa}(-\frac{\lambda_i}{\sigma^2} + \frac{1}{\sigma}\frac{\partial \lambda_i}{\partial \sigma})(1-S_i)$$

Finalmente, las derivadas de segundo orden de la log-verosimilitud del MCIM-GG están dadas por:

$$\frac{\partial^2 \mathbb{K}(\theta)}{\partial \omega_h \partial \omega_j} = \sum_{i=1}^{n_0} -\delta_i(1-\delta_i)\widetilde{x}_{ih}\widetilde{x}_{ij} + \sum_{i=n_0+1}^{n} -\delta_i(1-\delta_i)\widetilde{x}_{ih}\widetilde{x}_{ij}$$
$$+ \sum_{i=n_0+1}^{n} -\frac{1}{\kappa\sigma}(1-S_i)\delta_i(1-\delta_i)\widetilde{x}_{ih}\widetilde{x}_{ij} - \frac{1}{\sigma^2}S_i\delta_i^2\widetilde{x}_{ih}\widetilde{x}_{ij}$$

$$\frac{\partial^2 \mathbb{K}(\theta)}{\partial \beta_h \partial \beta_j} = \sum_{i=n_0+1}^{n} -\frac{1}{\sigma^2}S_i x_{ih}x_{ij}$$

$$\frac{\partial^2 \mathbb{K}(\theta)}{\partial \kappa^2} = \sum_{i=n_0+1}^{n} -\frac{1}{\kappa^2} + \frac{2}{\kappa^3\sigma}\ln y_i - \frac{4}{\kappa^6}\frac{\partial\psi(1/\kappa^2)}{\partial(1/\kappa^2)} - \frac{6}{\kappa^4}\psi(1/\kappa^2) + \frac{4}{\kappa^3}\frac{\partial S_i}{\partial\kappa} + \frac{6}{\kappa^4}(1-S_i)$$
$$- \frac{1}{\kappa^2}\frac{\partial^2 S_i}{\partial\kappa^2} - \frac{1}{\kappa^2\sigma^2}(\kappa\frac{\partial\lambda_i}{\partial\kappa} + \lambda_i)^2 + \frac{1}{\kappa^2\sigma^2}(\kappa\frac{\partial\lambda_i}{\partial\kappa} + \lambda_i)(-1 + \kappa\frac{\partial\lambda_i}{\partial\kappa} + \lambda_i + \frac{4\sigma}{\kappa})$$
$$- \frac{1}{\kappa\sigma}\frac{\partial^2\lambda_i}{\partial\kappa^2} - \frac{2}{\kappa^2\sigma}\frac{\partial\lambda_i}{\partial\kappa} + \frac{1}{\kappa^4}(-6\frac{\kappa\lambda_i}{\sigma} + 4 - 6\ln\kappa^2)$$

$$\frac{\partial^2 \mathbb{K}(\theta)}{\partial \sigma^2} = \sum_{i=n_0+1}^{n} \frac{1}{\sigma^2} + \frac{1}{\kappa\sigma^2}\ln y_i\frac{\partial S_i}{\partial\sigma} + \frac{2}{\kappa\sigma^3}\ln y_i(1-S_i)$$
$$+ (-\frac{\lambda_i}{\sigma^2} + \frac{1}{\sigma}\frac{\partial\lambda_i}{\partial\sigma})[\frac{1}{\kappa}\frac{\partial S_i}{\partial\sigma} + \frac{1}{\sigma}\frac{\partial\lambda_i}{\partial\sigma}(1-S_i) - \frac{\lambda_i}{\sigma^2}(1-S_i)] - (-\frac{\lambda_i}{\sigma^2} + \frac{1}{\sigma}\frac{\partial\lambda_i}{\partial\sigma})^2(1-S_i)$$
$$- \frac{1}{\kappa\sigma}\frac{\partial^2\lambda_i}{\partial\sigma^2}(1-S_i) + \frac{2}{\kappa\sigma^2}\frac{\partial\lambda_i}{\partial\sigma}(1-S_i) - \frac{2}{\kappa\sigma^3}\lambda_i(1-S_i)$$

$$\frac{\partial^2 \mathbb{K}(\theta)}{\partial \kappa\partial \sigma} = \sum_{i=n_0+1}^{n} \frac{1}{\kappa\sigma^2}\ln y_i\frac{\partial S_i}{\partial\kappa} + \frac{1}{\sigma^2\kappa^2}\ln y_i(1-S_i) + \frac{1}{\kappa}(-\frac{\lambda_i}{\sigma^2} + \frac{1}{\sigma}\frac{\partial\lambda_i}{\partial\sigma})\frac{\partial S_i}{\partial\kappa} - \frac{1}{\kappa\sigma}\frac{\partial^2\lambda_i}{\partial\kappa\partial\sigma}(1-S_i)$$
$$- \frac{1}{\kappa^2\sigma}\frac{\partial\lambda_i}{\partial\sigma}(1-S_i) + \frac{1}{\kappa^2\sigma^2}(\kappa\frac{\partial\lambda_i}{\partial\kappa} + \lambda_i)(1-S_i) + \frac{2}{\kappa^2}(-\frac{\lambda_i}{\sigma^2} + \frac{1}{\sigma}\frac{\partial\lambda_i}{\partial\sigma})(1-S_i)$$

$$\frac{\partial^2 \mathbb{K}(\theta)}{\partial \omega_h \partial \beta_j} = \sum_{i=n_0+1}^{n} \frac{1}{\kappa\sigma}\delta_i\frac{\partial S_i}{\partial\beta_h}\widetilde{x}_{ij}$$

$$\frac{\partial^2 \mathbb{K}(\theta)}{\partial \omega_h \partial \kappa} = \sum_{i=n_0+1}^{n} \frac{1}{\kappa^2\sigma}(1 + \kappa\frac{\partial S_i}{\partial\kappa} - S_i)\delta_i\widetilde{x}_{ij}$$

$$\frac{\partial^2 \mathbb{K}(\theta)}{\partial \omega_h \partial \sigma} = \sum_{i=n_0+1}^{n} \frac{1}{\kappa\sigma^2}(1 + \sigma\frac{\partial S_i}{\partial\sigma} - S_i)\delta_i\widetilde{x}_{ij}$$

$$\frac{\partial^2 \mathbb{K}(\theta)}{\partial \beta_j \partial \kappa} = \sum_{i=n_0+1}^{n} \frac{1}{\kappa^2\sigma}(1 + \kappa\frac{\partial S_i}{\partial\kappa} - S_i)x_{ij}$$

$$\frac{\partial^2 \mathbb{K}(\theta)}{\partial \beta_j \partial \omega} = \sum_{i=n_0+1}^{n} \frac{1}{\kappa\sigma^2}(1 + \sigma\frac{\partial S_i}{\partial\sigma} - S_i)x_{ij}$$

Apéndice B

Código en Matlab: simulación y aplicación del MCIM

B.1. Function "gg":

Función de densidad de la distribución gamma generalizada.

```
function [gg_pdf] = gg(y,lambda,kappa,sigma)

% De acuerdo a la parametrización propuesta por Manning (2005):

nu = sign(kappa).*(log(y)-lambda)./sigma;
gg_pdf = (abs(kappa)^(-2/abs(kappa)^2))./(sigma.*y.*abs(kappa)^(-1)*gamma(1/abs(kappa)^2)).*...
exp(nu./abs(kappa)-exp(abs(kappa).*nu)/abs(kappa)^2);

end
```

B.2. Function "kfun_mci_gg":

Función de log-verosimilitud del MCIM-GG y sus vector gradiente y matriz hessiana.

```
function [f,grad,H] = kfun_mci_gg(theta,X1,X2,Y)

% Índices y tamaños:
k1 = size(X1,2);
k2 = size(X2,2);
index0 = find(Y==0);
indexpos = find(Y>0);

% Parámetros sin regresión:
kappa = theta(k1 + k2 + 1);
sigma = theta(k1 + k2 + 2);

% Parámetros con regresión:
pred1 = X1*theta(1:k1);
```

```
pred2 = X2*theta(k1+1:k1+k2);
pred1 = pred1.*(pred1<=32).*(pred1>=-32)-32*(pred1<-32)+32*(pred1>32);
pred2 = pred2.*(pred2<=32).*(pred2>=-32)-32*(pred2<-32)+32*(pred2>32);
delta = 1./( 1+exp(-pred1) );
gama = exp(pred2);
lambda = log(gama(indexpos)) - log(1-delta(indexpos)) - 2*sigma*log(kappa)/kappa ...
         + log(gamma(1/kappa.^2)) - log(gamma(1/kappa.^2+sigma/kappa));

% Función log-verosimilitud:
aux0 = sum(log(delta(index0)));
[gg_pdf] = gg(Y(indexpos), lambda, kappa, sigma);
gg_pdf(gg_pdf<0.3e-310) = 0.3e-310;
aux1 = sum(log( (1-delta(indexpos)).*gg_pdf ));
f = -(aux0+aux1);

if isnan(f)==1 % Si encuentra un NaN (not-a-number)

f = realmax;
grad = ones(k1+k2+2,1); % 2*kk+2 parámetros a estimar
H = ones(k1+k2+2,k1+k2+2);

else

if nargout > 1

% X ordenada (en primeras filas van los que corresponde a Y=0):
X1s = [X1(index0,:); X1(indexpos,:)];
%X2s = [X2(index0,:); X2(indexpos,:)];

% Funciones auxiliares:
% lambda:
dlambda_dk = -2*sigma*(1-log(kappa))/kappa^2 - 2*psi(1/kappa^2)/kappa^3 ...
+ ( 2/(kappa^3)+sigma/(kappa^2) )*psi(0,1/(kappa^2)+sigma/kappa);
dlambda_ds = -2*log(kappa)/kappa - psi(1/kappa^2+sigma/kappa)/kappa;
d2lambda_dkdk = 2*sigma/kappa^3 ...
+ 4*sigma*(1-log(kappa))/kappa^3 ...
+ 4*psi(1,1/kappa^2)/kappa^6 + 6*psi(1/kappa^2)/kappa^4 ...
- (2/kappa^3+sigma/kappa^2)^2.*psi(1,1/kappa^2+sigma/kappa) ...
- (6/kappa^4+2*sigma/kappa^3)*psi(1/kappa^2+sigma/kappa);
d2lambda_dsds = - psi(1,1/kappa^2+sigma/kappa)/kappa^2;
d2lambda_dkds = - 2*(1-log(kappa))/kappa^2 ...
+ (2/kappa^4+sigma/kappa^3)*psi(1,1/kappa^2+sigma/kappa) ...
+ psi(1/kappa^2+sigma/kappa)/kappa^2;
% eta:
eta = exp(-kappa.*lambda./sigma);
deta_dm = eta.*(-kappa/sigma);
deta_dk = eta.*(-1/sigma).*(kappa.*dlambda_dk + lambda);
%deta_ds = eta.*(-kappa).*(-lambda./sigma^2+dlambda_ds./sigma);
```

```
d2eta_dkdk = -kappa.*eta.*d2lambda_dkdk./sigma - kappa.*dlambda_dk.*deta_dk./sigma ...
- 2.*eta.*dlambda_dk./sigma - lambda.*deta_dk./sigma;
% S:
S = (Y(indexpos).^(kappa/sigma)).*eta;
dS_db = -kappa.*eta.*Y(indexpos).^(kappa/sigma)./sigma;
dS_dk = -(lambda+kappa.*dlambda_dk-log(Y(indexpos))).*S./sigma;
dS_ds = kappa.*(lambda-sigma.*dlambda_ds-log(Y(indexpos))).*S./sigma^2;
d2S_dkdk = Y(indexpos).^(kappa/sigma).*d2eta_dkdk + ...
log(Y(indexpos)).*Y(indexpos).^(kappa/sigma).*deta_dk./sigma ...
+ log(Y(indexpos)).*dS_dk./sigma;

% Derivadas 1 de lnFV respecto a cada omega:
auxw0 = [ 1-delta(index0); (-delta(indexpos)) - (1-S).*delta(indexpos)./(kappa*sigma) ];
gradw = (X1s')*(auxw0);

% Derivadas 1 de lnFV respecto a cada beta:
auxb0 = -(1-S)./(kappa*sigma);
gradb = (X2(indexpos,:)')*(auxb0);

% Derivada 1 de lnFV respecto al kappa:
gradk = sum( 1/kappa ...
- log(Y(indexpos))./(sigma.*kappa^2) + 2*psi(1/kappa^2)/kappa^3 ...
+ 2.*S./kappa^3 - dS_dk./kappa^2  ...
+ (-1/sigma).*(kappa.*dlambda_dk + lambda)./(kappa^2) - 2.*(-lambda./sigma)./(kappa^2) ...
- 2/kappa^3 + 2*log(kappa^2)/kappa^3 );

% Derivada 1 de lnFV respecto al sigma:
grads = sum( -1/sigma ...
- (log(Y(indexpos))).*( 1-S )./(kappa*sigma^2)...
- ( -lambda./sigma^2 + dlambda_ds./sigma ).*( 1-S )./kappa  );

% Vector gradiente:
grad = - [gradw; gradb; gradk; grads];

if nargout > 2

% Derivada 2 de lnFV respecto a omega:
auxww01 = -delta(index0).*(1-delta(index0));
auxww02 = -delta(indexpos).*(1-delta(indexpos));
auxww03 = (-1).*(1-S).*delta(indexpos).*(1-delta(indexpos))./(kappa.*sigma);
auxww04 = Y(indexpos).^(kappa/sigma).*deta_dm.*(delta(indexpos).^2)./(kappa.*sigma);
auxww0 = [auxww01 ; auxww02 + auxww03 + auxww04];
auxww0 = sparse(auxww0);
Hww = X1s'*diag(auxww0)*X1s;

% Derivada 2 de lnFV respecto a beta:
auxbb0 = Y(indexpos).^(kappa/sigma).*deta_dm./(kappa*sigma);
auxbb0 = sparse(auxbb0);
```

```
Hbb = X2(indexpos,:)'*diag(auxbb0)*X2(indexpos,:);

% Derivada 2 de lnFV respecto a kappa:
A = kappa*dlambda_dk+lambda;
Hkk = sum( -1/kappa^2 + 2.*log(Y(indexpos))./(kappa^3*sigma)...
- 4*psi(1,1/kappa^2)/kappa^6 - 6*psi(1/kappa^2)/kappa^4 ...
+ 4.*dS_dk./kappa^3 + 6.*(1-S)./kappa^4 ...
- d2S_dkdk./kappa^2 ...
- A.^2./(kappa^2*sigma^2) ...
+ A.*(1/(kappa^2*sigma^2)).*(kappa*dlambda_dk+lambda+4*sigma/kappa)...
- d2lambda_dkdk./(kappa*sigma) ...
- 2*dlambda_dk/(kappa^2*sigma) ...
+ ((-6*kappa/sigma).*lambda+4-6*log(kappa^2))./kappa^4 );

% Derivada 2 lnFV respecto a sigma:
B = -lambda./sigma^2 + dlambda_ds./sigma;
Hss = sum( 1/sigma^2 + log(Y(indexpos)).*dS_ds./(kappa*sigma^2) ...
+ 2.*log(Y(indexpos)).*(1-S)./(kappa*sigma^3) ...
+ B.*(dS_ds./kappa+(dlambda_ds.*(1-S))./sigma-lambda.*(1-S)./sigma^2) ...
- B.^2.*(1-S) ...
+ (-d2lambda_dsds+2*dlambda_ds/sigma-2.*lambda./sigma^2).*(1-S)./(kappa*sigma) );

% Derivada 2 de lnFV respecto a omega y beta:
auxbw0 = delta(indexpos).*dS_db./(sigma*kappa);
auxbw0 = sparse(auxbw0);
Hbw = X2(indexpos,:)'*diag(auxbw0)*X1(indexpos,:);

% Derivada 2 de lnFV respecto a omega y sigma:
auxsw0 = (1+sigma.*dS_ds-S).*delta(indexpos)./(kappa*sigma^2);
auxsw0 = sparse(auxsw0);
Hsw = X1(indexpos,:)'*auxsw0;

% Derivada 2 de lnFV respecto a omega y kappa:
auxkw0 = (1+kappa.*dS_dk-S).*delta(indexpos)./(kappa^2*sigma);
auxkw0 = sparse(auxkw0);
Hkw = X1(indexpos,:)'*auxkw0;

% Derivada 2 de lnFV respecto a beta y kappa:
auxkb0 = (1+kappa.*dS_dk-S)./(kappa^2*sigma);
auxkb0 = sparse(auxkb0);
Hkb = X2(indexpos,:)'*auxkb0;

% Derivada 2 de lnFV respecto a beta y sigma:
auxsb0 = (1+sigma.*dS_ds-S)./(kappa*sigma^2);
auxsb0 = sparse(auxsb0);
Hsb = X2(indexpos,:)'*auxsb0;

% Derivada 2 de lnFV respecto a sigma y kappa:
```

```
Hsk = sum( log(Y(indexpos)).*dS_dk./(kappa*sigma^2) ...
+ log(Y(indexpos)).*(1-S)./(kappa^2*sigma^2) ...
+ B.*dS_dk./kappa ...
- d2lambda_dkds.*(1-S)./(kappa*sigma) ...
- dlambda_ds.*(1-S)./(kappa^2*sigma) ...
+ A.*(1-S)./(kappa^2*sigma^2) ...
+ 2.*B.*(1-S)./kappa^2 );

% Matriz hessiana:
H = - [Hww  Hbw'  Hkw  Hsw; ...
       Hbw  Hbb  Hkb  Hsb; ...
       Hkw' Hkb' Hkk  Hsk; ...
       Hsw' Hsb' Hsk  Hss];


end
end
end
end
```

B.3. Function "simular_ggci":

Función para simular datos de la variable respuesta del MCIM-GG

```
function [Y] = simular_ggci(X,theta)

% Notas:
% Las covariables que afectan a delta y gamma son las mismas.
% El orden de theta es: theta = [omegas; betas; kappa; sigma]. Un vector 12x1.
% Recordar: cdf_ggci = delta+(1-delta)*cdf_gg
% "gamma()" es diferente a "gama".
% Con función enlace logit para delta=P(Y=0) y log para gama=E(Y).

% Tamaño de muestra (n) y nro coeficientes de reg (k):
[n,k] = size(X);

% Parámetros sin regresión:
kappa = theta(2*k+1);
sigma = theta(2*k+2);

% Parámetros con regresión:
pred = X * [theta(1:k) theta(k+1:end-2)];
delta = 1./( 1+exp( -pred(:,1) ) );
gama = exp( pred(:,2) );

% Simular variable respuesta:
```

```
Y = zeros(n,1);
unif = rand(n,1);  % unif -> cdf_ggci
h = (unif-delta)./(1-delta);  % h = cdf_gg
aux = (unif>=delta);  % Las obs. donde Y>0
lambda = log(gama) - log(1-delta) - 2*sigma*log(kappa)/kappa + ...
         log(gamma(1/kappa^2)) - log(gamma(1/kappa^2+sigma/kappa));
Y(aux==1) = ( gammaincinv(h(aux==1), 1/kappa^2, 'lower').^(sigma/kappa) ).* ...
            exp(lambda(aux==1)) .* kappa^(2*sigma/kappa);

end
```

B.4. Script "simulacion_generar":

Rutina para simular datos de las covariables y la variable respuesta del MCIM-GG

```
clear;

% Fijar los parámetros verdaderos de reg:
omega = [-1  3.1 -1.5  1.3;    % delta aprox a 0.10
         -2 -3.1  0.9 -1.8;    % delta aprox a 0.20
         -3 -1.2  1.5 -2.9];   % delta aprox a 0.40
beta  = [-1 -2.9  1.1 -1.1];
kappa = 0.5;
sigma = 0.7;

% Fijar tamaños de muestra:
n = [200 500 1000 3000];

% Fijar número de simulaciones:
S = 1000;

for pro = 1:size(omega,1)

thetaverd = [omega(pro,:)'; beta'; kappa; sigma];

for tam = 1:size(n,2)

% Simular covariables:
X0 = ones(n(tam),1);
X1 = rand(n(tam),1);
X2 = normrnd(3,1,n(tam),1);
X3 = binornd(1,0.5,n(tam),1);
X = [X0 X1 X2 X3];
k = size(X,2);

colix = (k*tam-3) + k*size(n,2)*(pro-1);
```

```
colfx = (k*tam-0) + k*size(n,2)*(pro-1);
AX( 1:n(tam), colix:colfx ) = X;

for sim = 1:S

% Simular variable respuesta:
[Y] = simular_ggci(X, thetaverd);
coly = sim+S*(tam-1)+S*size(n,2)*(pro-1);
AY( 1:n(tam), coly ) = Y;

end
end
end
```

B.5. Script "simulacion_estimar":

Rutina para estimar el MCIM-GG en base a datos simulados. Tiene como insumo a las matrices de datos simulados que genera la rutina "simulacion_generar".

```
% Notas:
% (1) Se usa bases de Y que presentan dist GGCI
% (2) Función enlace logit para delta=P(Y=0) y log para gama=E(Y)

% Fijar los parámetros verdaderos de reg:
omega = [-1  3.1 -1.5  1.3;    % delta aprox a 0.10
         -2 -3.1  0.9 -1.8;    % delta aprox a 0.20
         -3 -1.2  1.5 -2.9];   % delta aprox a 0.40
beta  = [-1 -2.9  1.1 -1.1];
kappa = 0.5;
sigma = 0.7;

% Fijar tamaños de muestra:
n = [200 500 1000 3000];

% Fijar número de simulaciones:
S = 1000;

A_thetae = cell(S,length(n)*size(omega,1));
A_cob = cell(S,length(n)*size(omega,1));
A = cell(S,9*length(n)*size(omega,1));
contador = 0;
z = norminv(1-0.05/2,0,1);

for iPro = 1:size(omega,1)

for iTam = 1:length(n)
```

```
% Covariables:
coli = (k*iTam-3) + k*length(n)*(iPro-1);
colf = (k*iTam-0) + k*length(n)*(iPro-1);
X = AX( 1:n(iTam), coli:colf );

for iSim = 1:S

% Variable respuesta:
Y = AY( 1:n(iTam), iSim + S*(iTam-1) + S*length(n)*(iPro-1) );

% Índices de obs. donde Y=0 y Y>0:
index0 = find(Y==0);
indexpos = find(Y>0);
Xpos = X(indexpos,:);
Ypos = Y(indexpos);
Ybin = Y;
Ybin(Ybin>0) = 1;
n0 = length(index0);
npos = n(iTam)-n0;

% Opciones de optimización:
options = optimoptions(@fmincon,'Algorithm','trust-region-reflective',...
'GradObj','on','Hessian','user-supplied',...
'TolX',1e-16,'Tolfun',1e-16,'MaxIter',5000,'MaxFunEvals',5000,'TolPCG',0.001,...
'Display','off');

% MCIM-GG. Establecer val iniciales de theta:
[omega0] = glmfit(X,1-Ybin,'binomial','link','logit','constant','off');
[beta0,dev,stats] = glmfit(Xpos,Ypos,'gamma','link','log','constant','off');
beta0 = beta0.*(beta0>=-100).*(beta0<=100) - 100.*(beta0<-100) + 100.*(beta0>100);
alpha0 = 1/stats.sfit; % sfit: parámetro de dispersión Fam Exp
alpha0 = alpha0*(alpha0>=0.001)*(alpha0<=100) + 0.001*(alpha0<0.001) + 100*(alpha0>100);
theta0 = [omega0; beta0; sqrt(1/alpha0); sqrt(1/alpha0)];

% MCIM-GG. Optimizar log-verosimilitud:
[thetae,fval,exit,~,~,grad,H] = fmincon('kfun_mci_gg',theta0,[],[],[],[],...
[-100*ones(1,2*k) 0.1 0.1],[100*ones(1,2*k) 50 50],...
[],options,X,X,Y);

linf = thetae - z.*sqrt(diag(inv(H)));
lsup = thetae + z.*sqrt(diag(inv(H)));
theta = [omega(iPro,:)'; beta'; kappa; sigma];
cob = (linf<=theta).*(lsup>=theta);

% Guardar resultados:
A_thetae( iSim, iTam+length(n)*(iPro-1) ) = {thetae'};
A_cob( iSim, iTam+length(n)*(iPro-1) ) = {cob'};
```

```
coli = (9*iTam-8)+9*length(n)*(iPro-1);
colf = (9*iTam-0)+9*length(n)*(iPro-1);
A( iSim, coli:colf ) = {theta0',thetae',-fval,exit,grad,H,linf,lsup,cob};

% Contar escenarios:
contador = contador + 1;
disp([contador iPro iTam iSim])

end
end
end

% Sesgo y RECM:
thetae_MC = mean(cell2mat(A_thetae));
thetaverd21 = repmat( [omega(1,:)'; beta'; kappa; sigma] ,length(n),1 );
thetaverd22 = repmat( [omega(2,:)'; beta'; kappa; sigma] ,length(n),1 );
thetaverd23 = repmat( [omega(3,:)'; beta'; kappa; sigma] ,length(n),1 );
thetaverd2 = [ thetaverd21' thetaverd22' thetaverd23' ];
sesgo = thetae_MC' - thetaverd2';
ecm   = var(cell2mat(A_thetae))' + sesgo.^2;
sesgo = reshape( sesgo, [10,length(n)*size(omega,1)] );
ecm   = reshape( ecm, [10,length(n)*size(omega,1)] );
recm  = sqrt(ecm);

% Cobertura al 95%:
cobertura = mean(cell2mat(A_cob));
cobertura = reshape( cobertura, [10,length(n)*size(omega,1)] );
cobertura = cobertura.*100;
```

B.6. Script "aplicacion":

Estimación del MCIM-GG en el estudio de aplicación.

```
% Limpiar:
clc;
clear;

% Lectura de datos:
load('data.mat')
format short g

% Variable respuesta:
Y = data.gasto_edu;
n = size(Y,1);
length(find(Y==0))
sum(ismissing(Y))
```

```
% Covariables:
% Para delta = P[Y=0]:
X1 = [ones(n,1) data.vivienda_ind data.consumo_ind data.mh_estudian];
k1 = size(X1,2);
sum(ismissing(X1))
% Para gamma = E[Y]:
X2 = [ones(n,1) data.vivienda_ind data.consumo_ind data.sexo ...
data.centroestudio data.educacion];
k2 = size(X2,2);
sum(ismissing(X2))

% Índices de observaciones donde Y=0 y Y>0:
index0 = find(Y==0);
indexpos = find(Y>0);
Ypos = Y(indexpos);
Ybin = Y;
Ybin(Ybin>0) = 1;
n0 = length(index0);
X1pos = X1(indexpos,:);
X2pos = X2(indexpos,:);

% Opciones de optimización:
options = optimoptions(@fmincon,'Algorithm','trust-region-reflective',...
'GradObj','on','Hessian','user-supplied',...
'TolX',1e-18,'Tolfun',1e-18,'MaxIter',5000,'MaxFunEvals',5000,'TolPCG',0.001,...
'Display','off');

% MCI-G. Establecer valores iniciales de theta:
[b,dev,stats] = glmfit(X2pos,Ypos,'gamma','link','log','constant','off');
alpha0 = 1/stats.sfit; % sfit: parámetro de dispersión Fam Exp
beta0  = b;
[omega0] = glmfit(X1,Ybin,'binomial','link','logit','constant','off');
theta0 = [omega0; beta0; alpha0];

% MCI-GG. Establecer valores iniciales de theta:
theta0 = [omega0; beta0; sqrt(1/alpha0); sqrt(1/alpha0)];

% MCI-GG. Optimizar log-verosimilitud:
[theta2,fval2,exitflag2,~,~,~,hessian2] = fmincon('kfun_mci_gg',theta0,[],[],[],[],...
[-100*ones(1,k1) -100*ones(1,k2) 0.1 0.1],[100*ones(1,k1) 100*ones(1,k2) 50 50],...
[],options,X1,X2,Y);

% MCI-GG. Criterios de información:
logver2 = -fval2;
p2 = k1 + k2 + 2;
AIC2 = 2*p2 - 2*logver2;
AICc2 = AIC2 + (2*p2*(p2+1))/(n-p2-1);
```

```
BIC2 = log(n)*p2 - 2*logver2;
gama2 = exp(X2*theta2(k1+1:k1+k2));
RMSE2 = sqrt(mean((Y - gama2).^2));

% MCI-GG. P-valores:
ee2 = sqrt(diag(inv(hessian2)));
zp2 = theta2./ee2;
pval2 = 2*cdf('Normal', -abs(zp2), zeros(p2,1), ones(p2,1));
```

Apéndice C

Código en SAS: aplicación del MDP

El siguiente código en SAS es usado para la estimación del MDP-GG en el estudio de aplicación. El código es una adaptación del mostrado en Smith y Preisser (2018) y está basado en la función "proc nlmixed", una función cuyo uso general es estimar modelos mixtos no lineales.

```
proc nlmixed data = work.sheet1
maxiter=5000 GCONV=1E-28 FCONV=1E-28 HESS;

/* Indicar valores iniciales de los parámetros */
parms
omega0 = 0.48997
omega1 = 2.031
omega2 = 2.8882
omega3 = 0.096084
beta0 = 0.81177
beta1 = 0.10517
beta2 = 1.2877
beta3 = 0.046492
beta4 = -1.3612
beta5 = 0.073612
kappa = 0.77104
sigma = 0.77104;

/* Ecuaciones de regresión */
pred1 = omega0 + omega1*vivienda_ind + omega2*consumo_ind + omega3*mh_estudian;
pred2 = beta0 + beta1*vivienda_ind + beta2*consumo_ind + beta3*sexo
        + beta4*centroestudio + beta5*educacion;
delta = exp(pred1)/(1+exp(pred1));
mu = exp(pred2);

/* Función de log verosimilitud */
lambda = pred2 - 2*sigma*log(kappa)/kappa
         + log(GAMMA(1/((kappa)**2)))
         - log(GAMMA(1/((kappa)**2) + sigma/kappa));
```

```
eta = exp(-(kappa*lambda)/sigma);
if gasto_edu = 0 then fval = log(delta);
else if gasto_edu > 0 then do;
fval = log(1-delta) + log(kappa) - log(sigma)
       + ((1/(sigma*kappa))-1)*log(gasto_edu)
       - log(GAMMA(1/((kappa)**2)))
       - (eta/((kappa)**2))*((gasto_edu)**(kappa/sigma))
       + (1/((kappa)**2))*log(eta)
       - (1/((kappa)**2))*log((kappa)**2);
end;
model gasto_edu ~ general(fval);

run;
```

Bibliografía

Bayes, C. L. y Valdivieso, L. H. (2016). A beta inflated mean regression model for fractional response variables, *Journal of Applied Statistics* **43(10)**: p. 1814–1830.

Boyden, J. (2014). Young lives: an international study of childhood poverty: Round 3, 2009., *Base de datos. UK Data Service. SN: 6853* . https://discover.ukdataservice.ac.uk/Catalogue/?sn=6853&type=Data%20catalogue.

Duan, N., Manning, W., Morris, C. y Newhouse, J. (1983). A comparison of alternative models for the demand for medical care, *Journal of Business and Economic Statistics* **1(2)**: p. 115–126.

Grade (2015). Diseño y métodos del estudio Niños del Milenio en el Perú, Cuarta ronda de encuestas en el Perú: Informe de resultados. http://www.ninosdelmilenio.org/wp-content/uploads/2015/03/Dise%C3%B1o-y-m%C3%A9todos-del-estudio-.pdf.

Manning, G. W., Basu, A. y Mullahy, J. (2005). Generalized modeling approaches to risk adjustment of skewed outcomes data, *Journal of Health Economics* **24**: p. 465–488.

Morrow, V. (2017). A Guide to Young Lives Research. https://www.younglives.org.uk/sites/www.younglives.org.uk/files/GuidetoYLResearch_0.pdf.

SAS Institute Inc. (2018). The NLMIXED Procedure. https://support.sas.com/documentation/cdl/en/statug/63033/HTML/default/viewer.htm#nlmixed_toc.htm.

Smith, V. y Preisser, J. S. (2018). A marginalized two-part model with heterogeneous variance for semicontinuous data, *Statistical Methods in Medical Research* .

Smith, V., Preisser, J. S., Neelon, B. y Maciejewski, M. L. (2014). A marginalized two-part model for semicontinuous data, *Statistics in Medicine* **33**: p. 4891–4903.

Stacy, E. y Mihram, G. (1965). Parameter estimation for a generalized gamma distribution, *Technometrics* **7**: p. 349–358.

Stacy, E. W. (1962). A generalization of the gamma distribution, *The Annals of Mathematical Statistics* **33**: p. 1187–1192.

Su, L., Tom, B. y Farewell, V. T. (2009). Bias in 2-part mixed models for longitudinal semicontinuous data, *Biostatistics* **10(2)**: p. 347–389.

The MathWorks, Inc. (2018). Optimization Toolbox Functions: fmincon. https://la.mathworks.com/help/optim/ug/fmincon.html.

Tong, E. N., Mues, C. y Thomas, L. (2013). A zero-adjusted gamma model for mortgage loan loss given default, *International Journal of Forecasting* **29**: p. 548–562.